Author's final version. Version of record available on SpringerLink at: 10.1007/s10606-020-09372-2

Coordinative Entities: Forms of Organizing in Data Intensive Science

Drew Paine[#] and Charlotte P. Lee[*]
[#]Data Science & Technology Department, Lawrence Berkeley National Laboratory
[*]Human Centered Design & Engineering, University of Washington
pained@lbl.gov, cplee@uw.edu

Abstract. Scientific collaboration is a long-standing subject of CSCW scholarship that typically focuses on the development and use of computing systems to facilitate research. The research presented in this article investigates the sociality of science by identifying and describing particular, common forms of organizing that researchers in four different scientific realms employ to conduct work in both local contexts and as part of distributed, global projects. This paper introduces five prototypical forms of organizing we categorize as *coordinative entities*: the Principal Group, Intermittent Exchange, Sustained Aggregation, Federation, and Facility Organization. Coordinative entities as a categorization help specify, articulate, compare, and trace overlapping and evolving arrangements scientists use to facilitate data intensive research. We use this typology to unpack complexities of data intensive scientific collaboration in four cases, showing how scientists invoke different coordinative entities across three types of research activities: data collection, processing, and analysis. Our contribution scrutinizes the sociality of scientific work to illustrate how these actors engage in relational work within and among diverse, dispersed forms of organizing across project, funding, and disciplinary boundaries.

Keywords: Articulation work, Coordinative entities, Coordinated actions, Cyberinfrastructure, Data intensive science, Data science, Human infrastructure, Infrastructure, Synergizing

1 Introduction

Data intensive scientific research is accomplished through the construction and maintenance of complex sociotechnical systems supported by varied funding paradigms that bring together diverse arrangements of individuals, groups, and organizations (Edwards et al., 2007; Kaltenbrunner 2017; Lee et al., 2006; G. Olson et al., 2008a). It is now commonplace for scientists to engage with instruments located in geographically remote locations in concert with colleagues distributed across universities and institutes around the world, forming and engaging with multimorphous human infrastructures (Lee et al., 2006). Foundational work in Science and Technology Studies underscores how the production of scientific knowledge is a social endeavor (Fujimura 1996; Knorr-Cetina 1999; Latour 1987; Latour and Woolgar 1986) and within Computer Supported Cooperative Work (CSCW) studies of scientific collaboration typically examine particular endeavors to create and sustain sociotechnical infrastructures and their resources (Jirotka et al., 2013; Jirotka et al., 2006; Ribes and Lee 2010).

These infrastructure projects, often referred to as cyberinfrastructure (CI) within the United States funding paradigm, particularly focus on scientific needs when facilitating access to remote instruments (Borgman 2015; Finholt 2002) and enabling the journeys of data among sites and

across disciplinary lines as research projects evolve (Bates et al., 2016; Leonelli 2016). The visions for these cyberinfrastructure projects, in essence, are what Bowker and Star (1999) characterize as boundary infrastructures, those that "do the work that is required to keep things moving along" by having sufficient "play" to allow for local variation combined with consistent structure at scale to be stable through a "differing constitution of information objects within the diverse communities of practice that share a given infrastructure" (p.314). Infrastructures enable the multiple communities of practice at hand to readily "pull out the kinds of information objects" they respectively require in spite of their differences. Simultaneously local work infrastructures must emerge to support specific work tasks and practices as opposed to the "simple and universal services provided by traditional infrastructures" (Hanseth and Lundberg 2001).

Research on infrastructures and infrastructuring leaves us with conceptualizations of the interconnected, relational nature of information objects, stakeholders, projects and so on; long noting the need to operate among different embedded forms of organizing (Star and Ruhleder 1996). Yet too often our endeavors are not directly inspecting the forms and arrangements of organizing that the people doing the work create as they work within and among multiple overlapping sociotechnical infrastructures. In Lee and Paine (2015) we argue that there is a need in CSCW for conceptual models that describe and inspect our models of work rather than leaving them left to "linger in the shadows, designed for but unarticulated" since the work of design when building systems is in fact implicitly modeling sociality. CSCW needs theory of and for the field drawing out insights across scales as the increasing variety of sociotechnical arrangements leaves the community struggling to keep up and in "dire need of conceptual grounding" as we plunge forward in an "exciting era of research" (p.179). With CSCW's work to support data intensive science and infrastructure projects we are at a point where we need to examine more closely the conceptualizations of scientific sociality, specifically the forms of organizing used in different coordinated actions, that researchers and their infrastructural endeavors are building with and for—often all too implicitly. This is essential as scientific endeavors require the efforts of many allied communities of practice in the conduct of particular coordinated actions (e.g. interdisciplinary efforts). With this paper we shift focus, moving our collective perspective from particular infrastructure projects or funding structures to scrutinize the sociality of scientific work. We turn our attention to examining the myriad forms of organizing that data intensive scientists rely upon and create as they grow and sustain diverse human infrastructures to conduct their work through different overlapping and intersecting coordinated actions (interdependent efforts of two or more actors who through their individual activities are working towards a particular goal through fields of work (Lee and Paine 2015, p. 184)). Rather than another example of infrastructuring or design, our empirical contribution is an articulation of these recurrent ways of organizing that appear across our four research sites which we characterize as *coordinative entities*.

Schmidt and Wagner (2004) posit that "contemporary cooperative work is generally characterized by heterogeneous and often widely ramified arrangements of actors immersed in complex interdependencies of varying scope, intensity, and degrees of coupling"; what we call coordinated actions. A key enabler of the visions underlying sociotechnical infrastructure projects and data intensive work has been the ability to foster, and the necessity of employing, new arrangements of stakeholders (Bietz et al., 2010; Lee et al., 2006). Clarke and Star (2008) describe the situation of infrastructures in and among social worlds by observing that "infrastructures can

be understood, in a sense, as frozen discourses that form avenues between social worlds" (p.115) and the database development studied by Bietz and Lee (2009) illustrates how a component of a nascent infrastructure unfolds as a boundary negotiating artifact (Lee 2007) on the path to freezing in place as a particular discourse, aligning and solidifying different social worlds. This is essential as accomplishing collaborative scientific work and the enactment of infrastructural components requires scientists craft "doable problems" by aligning across multiple scales or forms of organizing (Fujimura 1987, 1996).

Elucidating more about the ways people craft these avenues and frozen discourses while attempting to arrive at doable problems is a complex task when data intensive research practices are constantly evolving. This requires our continued focus and this paper's conceptual contribution is a step towards identifying arrangements scientists rely upon while working within and among varying social worlds through myriad different coordinated actions. Building upon the human infrastructure (Lee et al., 2006) and synergizing notions (Bietz et al., 2010) our study traces the webs of collaboration of four scientific groups, from four different disciplines. Furthermore, we investigate how these groups create and draw upon these webs as they work across project, infrastructure, institutional and organizational, and disciplinary boundaries to conduct scientific research. By following what has been called the careers (Harper 2000) or journeys (Bates et al., 2016; Leonelli 2016) of software, data, instruments, and so on, we develop a new lens for understanding aspects to the diverse ecologies of work fundamental to accomplishing contemporary data intensive science. We answer the question: *How are scientists (re)organizing when conducting data intensive work over time?*

2 Background

CSCW investigations of scientific work span early research on collaboration and collaboratories to more recent studies of cyberinfrastructure and the myriad social and technical elements of such projects. Here we examine key points in the trajectory of this work from teams and tools to infrastructure projects to highlight a gap in our understanding of the social arrangements and forms of organizing in collaborative scientific work that our study addresses.

2.1 Teams, Projects, Tools, and Scientific Collaboration

Investigations of collaborative scientific work in CSCW reach back to the field's formation. This work tended to not focus on the dynamics of ongoing research among and across diverse organizational boundaries as our work is doing. This work was often tool and/or team centric, often in service of particular projects or organizational endeavors.

Kraut et al. (1988, 1986) study collaborative pairs and posit that scientific relationships form to combine material and intellectual resources to be able to do work and that proximity influences researchers ability and willingness to work together—foreshadowing Olson and Olson's (2000) point that distance matters. Kraut et al. (1988) offer an early typology of tools to support scientific collaborations (communication tools, coordination and management tools, and task-oriented tools). Contemporaneously, Latour and colleagues' (Latour 1987; Latour and Woolgar 1986)

classic studies of laboratories characterized the relations among different elements in scientific work, asserting that social worlds exist in relationships between human and non-human "actants," reshaping the way ethnographers explored the work of teams in laboratories. Chompalov and Shrum (1999) in turn investigated the formation of teams of scientists from multiple organizations to demonstrate how technological practice serves as a predictor for success in routines and in part to advocate for studying more than just singular laboratories, which CSCW collaboratory studies accomplished too.

Subsequently, CSCW scholars studied collaboratories as systems built to facilitate cooperative scientific work independent of location or time (Finholt 2002; Wulf 1993). J. Olson et al. (2008b) synthesize a theory of remote scientific collaboration from studies of collaboratories to offer suggestions of what makes for successful collaboratory projects and tools. Bos et al. (2008) use the same dataset to offer a taxonomy of seven types of collaboratories, exploring key technology organizational issues so as to identify organizational patterns to support funders and project managers creating new projects. Similarly, a 2015 United States National Research Council report examines how to improve the effectiveness of "team science" and offers a taxonomy of this type of work defining seven dimensions (National Research Council 2015). With dimensions such as "diversity of team or group membership," "team or group size," and "permeable team and organizational boundaries" (p.26) this taxonomy is again focused on characterizing particular forms of collaborative scientific projects and not how science unfolds in day-to-day contexts. These are useful perspectives which provide insights with which to shape particular tools or to influence the structure of projects. But more is needed to clarify and describe day-to-day arrangements PIs employ to accomplish their research across project and organizational boundaries over time.

Following investigations of collaboratories and teams, a vast body of CSCW work has emerged studying particular projects building infrastructures for scientific collaboration, in the US commonly characterized as cyberinfrastructure and in Europe as e-Research (Jirotka et al., 2013; Ribes and Lee 2010). Cyberinfrastructure studies draw upon Star and Ruhleder's (1996) relational infrastructure orientation as part of the work of characterizing evolving infrastructuring projects (Borgman 2007; Edwards 2010; Edwards et al., 2007). This body of work examines issues of design and development across disciplinary boundaries and time scales in particular projects, while better characterizing the impediments and opportunities to sharing scientific resources such as data and software across domains with varying cultures.

2.2 Beyond Individual Cyberinfrastructure Projects

Studies of cyberinfrastructure development continue to usefully inform our understanding of project dynamics. However, the historic tendency of CSCW (Ribes and Lee 2010) to focus on the development of individual infrastructure projects leaves missing, at least in part, richer understandings of the ways in which day-to-day scientific work, including tool and system development, are achieved. Going forward, instead of focusing on particular infrastructure projects and the funding structures that enable them we need to examine the different organizational arrangements scientists form and engage in as they seek to discover the limitations and opportunities of their data, method, theory, instruments, and projects. Individual scientists and

their groups do not participate solely in relatively well defined projects. They function entrepreneurially, working across project boundaries. We understand this in part due to cyberinfrastructure work that examines and emphasizes the multiple, overlapping roles of stakeholders in human infrastructures who engage in synergizing work as we examine in the next section.

Science-oriented infrastructure studies have described numerous sociotechnical challenges inherent in developing, growing, and sustaining distinct infrastructure projects as research interests and funding prerogatives change (Bietz et al., 2012; Karasti et al., 2010; Ribes 2014, 2017; Steinhardt and Jackson 2014). Designing, building, and sustaining such distinct endeavors surfaces complex social dynamics and tensions where domain scientists goals need to align with computer science research interests (Ribes and Finholt 2009). Projects have to determine how decision making should be accomplished, whether top down by management or bottom up by individual researchers, all while managing information exchange across sites of work (Lawrence 2006). Infrastructuring endeavors must also navigate the overarching funding structures that instantiate the conditions for work and enable diverse stakeholders to come to the table and be successful over time across funding streams with varying policy demands (Kaltenbrunner 2017; Kee and Browning 2010). The temporal scales of these projects, both planned and actual, also change the dynamics of researchers work; design goals and actions shift when building for decadal time spans rather than the short term alone (Cohn 2016; Ribes and Finholt 2009).

Challenges that are often outside the bounds of designated infrastructure development efforts include the facilitation of the sharing of the products of these efforts, namely software and data (Birnholtz and Bietz 2003). Cyberinfrastructure studies underscore the recurring issue of trust scientists face when sharing and reusing data (Edwards et al., 2011; Faniel and Jacobsen 2010; Zimmerman 2008). Scientists reusing data must be able to determine the appropriate questions to ask of this product to understand the contexts of its production and how they may appropriately employ this product for new questions (Rolland and Lee 2013). There is a need to examine and explain different cultures of work around data as behaviors of openness and secrecy vary among different fields (Velden 2013) and even within projects of one organization in the same broad discipline (Vertesi and Dourish 2011). Sharing scientific software raises similar issues as the incentives to make complex custom work available to others vary based on community cultures and the requirements of funders and publishers (Howison and Herbsleb 2011).

Many of the pressing questions that have been surfaced by the study of infrastructure development and use are questions that we can also ask more broadly of scientific collaboration in general. With the advent of new technologies more kinds of science are becoming simultaneously increasingly collaborative and data-intensive. In data-intensive science most collaborations are not infrastructure development efforts per se, however, they do usually overlap, plug into, draw on, and crisscross infrastructure in very interesting ways. This paper, however, is focused not on infrastructure. This paper is focused on the interesting ways that scientists organize themselves in order to accomplish science, specifically the scientific steps of: data collection, processing, and analysis. These notions have a great deal of resonance with earlier work on *human infrastructure* but here we look at human infrastructure apart from the context of cyberinfrastructure.

2.3 Human Infrastructure, Synergizing Work, and Gaps in Understanding Forms of Organizing

Scientific infrastructure studies in CSCW surfaces some organizational forms that emerge as work is completed over time. Theorizing about the human infrastructure of cyberinfrastructure as a lens has been useful for clarifying that multiple forms of organizing are necessary in order for diverse stakeholders to work together to develop infrastructural components and that these forms of organizing may occur in parallel, forming and dissolving as needed (Berman 2001; Lee et al., 2006). The follow-on concept of synergizing (Bietz et al., 2010) further unpacks how different actors come together and foster different relationships in order to enable new work to happen. These insights help us to frame a gap in our understanding: that of identifying key regular forms or arrangements of organizing, necessary for getting data intensive science done in the face of constantly changing technologies, organizational contexts, and scientific knowledge.

The human infrastructure of cyberinfrastructure has come to refer to not only the necessity for complex endeavors to rely and draw upon a variety of collaborative forms over time and often simultaneously (e.g. groups, teams, networks, organizations, etc.), but also the complex interactions between the activities of networks, place-based organizations, groups, and consortia (Bietz et al., 2010). These collaborative structures configure, and are configured by, infrastructure creation—they are enacting different coordinated actions to address myriad overlapping, shifting common fields of work. Collaborative work on such resources can both draw upon and contribute to various infrastructures and infrastructural resources (components of infrastructure such as software, data, practices, etc.) and have very different characteristics (e.g. networks vs. teams) which may be invoked serially and in parallel (Lee et al., 2006). Human infrastructure posits that participation takes many forms and that no one type such as teams, networks, or organizations can account for the whole—human infrastructure is complex and heterogeneous. Participation may take some or all of these forms simultaneously, and participants in a CI may not even be fully aware of the breadth of forms they are involved in at any given time.

Bietz et al. (2010) defined the notion of synergizing as a social process that is fundamental to scientific cyberinfrastructure work. The concept of synergizing was established to refer to the work necessary to enact productive infrastructural relationships. It is a concept where strategic coordinated actions are undertaken "in pursuit of greater combined effects than individuals, groups, or organizations could effect on their own" since synergy can emerge from bringing different individuals or collaborative arrangements, or resources, together in a productive relationship as an entity rather than "from scratch" (Bietz et al., 2010, p. 252). Synergizing understands the embeddedness and relational structure of cyberinfrastructure as both a development goal and as a resource for development work. Synergizing includes two key development activities: leveraging and aligning. Leveraging refers to using existing relationships as a resource to create or maintain existing relationships. Aligning is the work required to make an infrastructural relationship productive by ensuring that there is compatibility among components, it is essential to ensuring that a common field of work for a cooperative work arrangement exists.

Building upon synergizing, subsequent work in healthcare infrastructure development characterized the potential opposite case of reverse synergy which can directly lead to breakdown of productive relationships and an instantiated infrastructure (Langhoff et al., 2018). Reverse

synergy happens when the effort individuals need to exert to align diverse coordinated actions "creates enough cracks in the inertia in a given information infrastructure, erodes enough social capital, or in other ways require an amount of alignment work, articulation work, or other types of coordination" (p.51) to be greater than the potential benefits of successfully synergizing. This cautionary tale reminds us that there are potential downsides to the process of synergizing where fields of work already stably exist. No matter what, in collaborative work diverse sets of relationships must be built among diverse sets of entities (whether technologies, people, organizations, communities, or so on) that will vary depending on context. Sometimes this will be productive and positive, in others counterproductive and negative.

Schmidt (1990) developed a broad framework for analyzing cooperative work and the notion of cooperative work arrangements to describe entities CSCW scholars investigate. He describes cooperative work as "constituted by work processes that are related as to content, that is, processes pertaining to the production of a particular product or type of products" (p.10). Cooperative work arrangements emerge when multiple individuals with diverging interests and motives come together to complete a task (p.12). A given cooperative work arrangement exists in relation to a particular common field of work that is creating particular products. In the context of scientific collaboration and infrastructure this could be any number of products, from datasets or data analysis software to particular organizational policies. Schmidt continues to explain that collaborative work requires an "organizational form" where "an organization is conceived as a stable pattern of cooperative relations" (p.38).

A challenge we face in CSCW today is more consistently identifying and understanding common forms of organizing through which we can identify synergizing and reverse synergizing. We need a conceptual lens that helps us identify and trace out these heterogenous coordinated actions. While Bietz et al. (2010) were primarily concerned with understanding how diverse entities come together to be productive and was not "overly concerned with creating comprehensive lists of relationships and entities" we find that there is a need to identify and to undertake listing of some common forms of organizing that emerge. Our work here is concerned with types of entities, with the understanding that this is not a comprehensive list, and tracing some of their relationships in the context of data intensive science. We return to this discussion about common forms and link it with a discussion of explorations of meso-level theory in the Discussion.

3 Research Sites & Methods

Our qualitative study was designed to investigate the work of four different scientific groups and their webs of work. We identified four Principal Investigators (PIs) and members of their research groups including graduate students, postdoctoral researchers, and research scientists at the University of Washington in Seattle, WA engaging in data intensive work with a variety of collaborators. We intentionally began our inquiry with PIs and their groups instead of a particular cyberinfrastructure project so that we can follow the journeys and trajectories of work outward from a clearly bounded starting point. We chose PIs at a university as a starting point since they are a key way funding is distributed in the United States and elsewhere, setting research programs, fostering the careers of other researchers and students, and distributing and managing resources

(cf. Knorr-Cetina 1999; Latour and Woolgar 1986). In order to get closer to the actual work of creating and sustaining collaborations when doing data intensive research among different contexts, we interviewed not only PIs but also their graduate students and, when applicable, postdocs, research scientists, and undergraduate researchers. *The names of individuals, groups, and projects are referenced here using pseudonyms to protect our informants' privacy as much as possible*.

3.1 Research Sites: Four Principal Investigator's Research Groups

All of our research sites were chosen for their self-identified data and software intensive research and willingness to participate in a longitudinal study—further details about our initial sampling is available in (Paine et al., 2014). All four groups are engaged in multiple research projects and where possible, we aimed to follow two active projects that were identified with the help of each PI to make our investigation feasible given our own resource limitations. Due to the variation among the work of the groups the number of projects being studied does vary.

3.1.1 Hank: Climate Science Modeling

Hank is an atmospheric scientist studying the interaction of different Earth processes that shape the global climate cycle. Hank's group had four Doctoral students (Anita, Bryan, Dane, and Palmer) during our study. Each PhD student was working on individual dissertation projects examining the effect of low-frequency changes of different variables on the sensitivity of climate models.

3.1.2 Waldo: Marine Geophysics

Waldo is a marine geophysicist studying submarine volcanoes and mid-ocean ridge hydrothermal systems to better understand how the Earth's physical structure is changing. Waldo's research group was composed of three Doctoral students (Dahl, Rollin, and Megan) working on two underseas seismology projects using ocean cruises to collect data to input in computational models.

3.1.3 Martin: HIV Microbiology

Martin is a virologist studying the Human Immunodeficiency Virus (HIV) with wet lab and computational biological research to examine the efficacy of vaccines and the evolution of different strains of HIV in the search for effective treatments. Martin's research group is composed of multiple doctoral and undergraduate students, postdoctoral researchers, and a large research scientist staff. Our inquiry focused on two projects where the group used pyrosequencing techniques that required the development of new wet lab molecular techniques and subsequently new data processing and analysis software and practices.

3.1.4 Magnus: Cosmology with Radio Telescope Arrays

Magnus is an observational cosmologist studying a period of the Universe's development known as the Epoch of Reionization (EoR) using novel radio telescopes. Magnus's research group was composed of three postdoctoral researchers (Brianna, Igor, and Jonah) with three doctoral students

(Abner, Nima, and Peg) and a rotating cast of undergraduate students. They were primarily focused on the development of data analysis software for use with the Widefield Radio Telescope (WRT), an instrument producing petabytes of data for analysis that requires new software analysis approaches (Paine 2016; Paine and Lee 2014, 2017).

3.2 Research Methods: Data Collection and Analysis

Our qualitative study relies upon three forms of iterative data collection and analysis that took place over 2011–2015. We collected almost 40 hours of interviews and 47 hours of observation among the four sites over repeated episodes. This inquiry was intentionally oriented around the work of a PI and their laboratory or group since these bounded entities are a key source of synergizing activity. Magnus's group was the focus of deeper ethnographic inquiries as the first author conducted dissertation work. We conducted semi-structured interviews with the PIs and researchers in each group to have them walk us through their research work, when possible attended meetings of the groups to learn about ongoing tasks, and collected and analyzed artifacts from the groups to augment our observations and interviews. Artifacts ranged from publications to internal Wikis, email threads, software and data repositories, and public websites. The findings in this paper are derived from analysis of the interviews with contextual details for the cases filled in with our other sources of data. Our semi-structured interviews took place over multiple rounds. Each interview was recorded and professionally transcribed and cleaned by the member of the research team who conducted it.

The first round of interviews took place with PIs so that we could learn about their research. The four interviews were scheduled for around an hour in Spring 2011 and ranged between 52 and 82 min (avg. 66 min)—the four discussed in this paper are a subset of 20 PIs who we interviewed before enrolling this subset in our longitudinal study. The second and third rounds with members of each group (students, post-docs, research scientists) drew out information about the work these individuals take part in as part of a PI's group. The second round of interviews took place in Spring 2013, ranged from 25 to 78 min (avg. 47 min), and were conducted with: ten members of Martin's group, four members of Magnus's group, three members of Waldo's group, and five members of Hank's group. The third round was conducted in Winter 2014, ranged from 55 to 125 min (avg. 73 min) and were conducted with four members of Martin's group, four members of Magnus's group, two members of Waldo's group, and four members of Hank's group. Eleven of the thirteen individuals interviewed in round three were previously interviewed in round two.

The PI interviews (round one) and first interviews with research group members (round two) were designed to provide us with a baseline understanding of each research group's science, different projects, types and sources of data, software being created and used, and the collaborators they work with. We followed the trajectories or journeys of resources (Harper 1997) such as data and software across and among different research groups and their collaborators over time to trace out arrangements of relationships necessary to achieving scientific goals and enacting infrastructures as they work to create, access, use, and share a variety of resources.

Our initial open coding analysis of these interviews began to surface what we came to describe as different "entities" involved in the collaborative work of each group. For our third round of interviews we re-interviewed group members to have them walk us through their work in-depth.

This interview protocol specifically asked interviewees about their work collecting or producing data, processing data, analyzing data, sharing data, and archiving data as well as who was involved in each activity. This guided our focused coding of all of the interviews to surface the entities we were seeing and to guide axial coding to elicit the different characteristics of each entity.

Analysis of this data took place over multiple iterations to guide our ongoing data collection. The round one and two interviews were closed coded for the questions in our protocol and open coded for emergent themes about each individual and group's work (Charmaz 2014; Emerson et al., 1995; Weiss 1995). Using this coding our research team wrote memos on, and created diagrams of, each group's projects and work to draw out themes from which we began to see the different entities in our typology emerging. As we analyzed this data we found it difficult to disentangle the complexity of the organizations of these collaborations and their work with data. This difficulty as well as literature reviews led us to decompose the activities our interviewees were engaging in as well as to try to categorize the different relationships they form to do this scientific work. Through this analysis we began to develop distinctions between a local research group, intellectually close collaborators, and collaborators whose work is more diffused from our interviewee's day-to-day concerns yet still connected.

This initial analysis and initial categorization and decomposition of our subject's work informed the design of our round three interview protocol. This protocol was designed to not only explicitly focus on different research activities but also attempted to further draw out the involvement of different entities throughout this work based on our initial categorizations. We open coded these interviews and wrote memos comparing each group's project work to our prior analysis to help us better define the different types of entities we were seeing emerge in our data. To re-assess and triangulate our initial findings about these entities, the first author and two members of our research team then re-coded all of the round one, two, and three interviews using this emerging framework of coordinative entities as our code book in a round of focused coding. New memos were written to describe how each research group collaborates to accomplish different research activities over time. These final memos informed the cases presented in this paper and our prior publications about cosmology software development (Paine and Lee 2014, 2017) and data processing work (Paine et al., 2015). From this analysis we have developed this typology of coordinative entities.

Due to the breadth of the work undertaken, we are limited in our ability to assess when such differences arise from variations in disciplinary practices. We are also not yet able to richly describe the dynamics and evolution of these collaborations as this initial study was not ethnographic. Our first step is necessary to begin to develop a language and framework to guide our desired, longer-term ethnographic inquiries.

4 Five Types of Coordinative Entities

Building on previous research on human infrastructure and synergizing (Bietz et al., 2010; Lee et al., 2006), we have developed the notion of forms of organizing we call coordinative entities since data intensive science can often be organizationally intensive. Just as the process of developing infrastructures for science requires synergizing, so too does the conduct of science over the course of various research activities (data collection, processing, and analysis), institutions and funding

structures, and disciplinary boundaries. We define five prototypical coordinative entities, particular forms of organizing, that appear and reappear in our empirical data; although we anticipate that other coordinative entities exist and that studies by us or others would uncover more. Our goal is not to exhaustively list entities or characteristics, but rather to put forth a conceptual lens that can frame, seed, motivate, and contextualize further inquiries into understanding, designing, and developing for complex collaborative work in science and beyond. Our analysis identified three variable characteristics for each coordinative entity: *organizing focus*, *formality of organization*, and *planned permanence*.

4.1 Three Characteristics that Define a Coordinative Entity

Here we draw out three characteristics that define the coordinative entities that are created or engaged with as forms of organizing in the scientific work we studied. These characteristics emerge from our axial coding of our qualitative data. Through future work we expect that additional characteristics may be derived, and the three existing characteristics further developed.

Organizing Focus. Organizing focus is the combined scoping and motivation for creating an organizational form. Each coordinative entity's focus and reason for existence can vary widely, sometimes with multiple connected reasons. We currently find three common focii in our data: 1) advancing a particular research program; 2) to complete specific intermediary tasks; and 3) to provision and sustain resources. Example reasons include: a Principal Investigator wants to advance their agenda; a funding agency or community body wants to create and sustain particular resources for diverse scientific goals; individuals need to share a resource to accomplish a task.

Formality of Organization. Formality of organization refers to how binding agreements are constructed and used in a coordinative entity, the particular organizing of relations administering and directing a given coordinative entity. An obligation is imposed on an individual or entity to do some task by means of an agreement (whether formalized or not). This ranges from no formal structure for nascent, ad-hoc arrangements to formal with codified rules and regulations governing the entity's existence. Can be structured with a commercial or monetary contract (e.g. Co-PIs or services for hire) or informal but potentially detrimental to professional reputation or relationships. Example formalities of organization include: none; formal where individuals join a group as researchers through a defined process and must abide by a set of rules; informal where individuals from one entity come together with those from another entity to accomplish a task.

Planned Permanence. Planned permanence is the intended permanence of a given coordinative entity to exist over time. Planned and not known "because it often cannot be predicted how long" a given entity may last (Lee and Paine 2015, p. 186). Regardless of whether the entity is temporary or permanent, shared practices, artifacts, and terms need to be created. Examples include: short-term to access resources (e.g. datasets, instruments, software); long-term to sustain a research agenda; variable-term until a task is completed.

4.2 Five Types of Coordinative Entities

There are five coordinative entities in our typology today, Fig. 1 (summarized in Table 1).

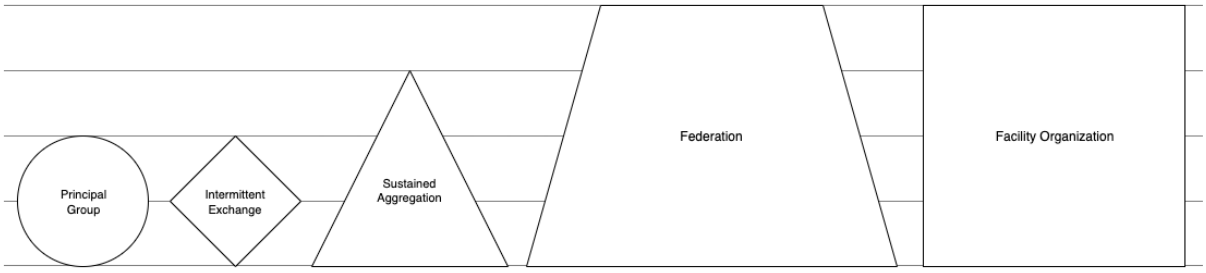

Figure 1. Visual representation of each type of coordinative entity.

Principal Group (PG). The Principal Group entity is the organizing of a particular scientific Principal Investigator (PI) and the members of that group. Principal Investigators are managers who have control over monetary and human resources with significant autonomy in an organization. The organizing focus of a PG is to advance the PI's research agenda, even as all members of the PG have agency and individual goals that align and diverge with the PI's. Individuals may include undergraduate, masters, and doctoral students; postdoctoral researchers; research scientists; research staff; and laboratory managers. The Principal Group's formality of organization is defined by the PI organizing and running the group (and often delegating responsibilities), determining which individuals join and under what conditions. The planned permanence of a PG is long-term with the intention of enduring over time so long as the PI sustains and renews the group.

Intermittent Exchange (IE). An Intermittent Exchange's organizing focus is to work towards the completion of a specific task by creating a common field of work between members of a PG and individuals from other organizations. IEs have no formality of organization because this type of entity emerges to accomplish a task. This entity's planned permanence is variable-term. The intention may be to last for a short period of time yet end up being carried on for a long period of time if a task or common field of work are complex.

Sustained Aggregation (SA). A Sustained Aggregation's organizing focus is to bring together at least two Principal Groups (along with their resources) to address a common, potentially evolving, research problem by creating and sustaining the necessary common fields of work. The formality of organization of SAs can span a wide spectrum depending on the wishes of the PGs. They may have a formal organization if grant proposals are co-authored with multiple investigators or there may simply be an informal agreement among PIs choosing to pool resources to tackle a shared problem of interest. SAs are planned and run without necessarily formally codifying rules for an administrative structure. A given SA's planned permanence is variable-term, existing so long as the PG's wish to continue work on a shared research problem and maintain the relationship.

Federation. A Federation's organizing focus is to create scientific resources (instruments, datasets, software, etc.) for a set of scientific pursuits specified through a charter. The charter provides a written agreement that is contractual in nature in that it articulates arrangements but need not be a legal document or legally enforceable. Federations are formally organized, with a board or management group tasked with fulfilling the charter's intellectual aims by setting forth

rules and guidelines for membership and participation. A Federation's planned permanence is long-term so that the resources it creates can be sustained over time to address the scientific aims identified. In practice a Federation may only exist for a short or medium term. Many of the cyberinfrastructure projects studied by CSCW researchers would be categorized as Federations.

Table 1. The five coordinative entities and their characteristics.

Characteristics of each entity	*Principal Group (PG)*	*Intermittent Exchange (IE)*	*Sustained Aggregation (SA)*	*Federation*	*Facility Organization (FO)*
Organizing Focus	Created by PI to further their research agenda	Enacted by PG member towards completion of a specific task	Enacted by PGs to work together to address a shared research problem	Organization created with a charter to create and use resources for specified intellectual pursuits	Organization created with a charter to sustain specified information, technical, and human resources for (re)use by diverse stakeholders
Formality of Organization	PI organizes and runs the group, deciding which people join and under what terms	Nothing formalized	PGs plan, organize, and run without necessarily codifying rules for the SA	Board or management group specifies and enacts charter, codifying rules declaring who may join and for participation	Board or management group fulfills a charter's aims to make resources available to a scientific community, often without membership requirements
Planned Permanence	Long-term, enduring so long as the PI sustains & renews the group	Variable-term, existing until a specified task is completed	Variable-term, sustained so long as the given contractual relationship is maintained	Long-term to sustain resources for those working on specified intellectual pursuits	Long-term to sustain resources and ensure their wide availability

Facility Organization (FO). A Facility Organization's organizing focus is to sustain information, technical, and human resources specified through a charter. FOs are not organized to sustain particular intellectual aims, they sustain resources produced by other entities so that they may be used or reused by such parties who have particular intellectual purposes in mind—a key difference from Federations. Again, here a formal charter indicates a written agreement that is contractual in nature in that it articulates arrangements but need not be a legal document or legally enforceable. The formality of organization for FOs also results in a board or management group taking responsibility for fulfilling the charter's aims and providing resources, often without

membership requirements. FOs may provide resources (such as software or datasets) freely to anyone via an open system. They may also provide resources as a service in exchange for currency (e.g. genome sequencing firms offering sequencing as a service). FOs have long-term planned permanence to ensure the wide availability of the entity's resources over time.

In our findings section we will use these entities to discuss work in our four research sites for three types of research activities. We will then reflect on these entities and our cases in the discussion.

5 Examining Four Cases

Following the threads of scientific data among individuals, data, and software we see these five entities employed across our four different research sites. We draw attention here to the ways work producing, processing, and analyzing data is accomplished through these forms of organizing so that different stakeholders, scientific problems, and resources can be aligned and productively combined in the course of this collaborative work.

5.1 Hank: Climate Modeling Using Facility Organization Resources

The threads of climate research we followed in Hank's Principal Group offers a baseline of scientific work with relatively (especially if compared to Magnus's PG described later) non-complex forms of organizing across the multiple key research processes. Hank's PhD students complete individual projects by utilizing Facility Organization climate models and datasets from FOs and obtaining advice from varying IEs, Fig. 2. Students in Hank's PG study a variety of different phenomena by iterating these existing climate models—complex software assemblages of theory, data, and past executions of the model (cf. Edwards 2010)—through adjustments of variables and the integration of different datasets to test new hypotheses. This PGoriented work contrasts to the often more organizationally complex work of the FOs that are creating and sustaining climate models. Instead of working with dozens of people, e.g. doing "very large collaborations" Dane a PhD student noted, members of this PG may interact with two or three others through Intermittent Exchanges when completing tasks. This case demonstrates work where entities are invoked with short-term planned permanence.

> Our group doesn't do very large collaborations. Some people do very, very large collaborations like 20 people or so, typically [Hank's] group, only work with like two or three people. A lot of people will just go through their entire graduate degree with basically two author papers with [Hank]. (Dane, PhD student).

The processes of Hank's PG's **collecting data** and climate models most commonly relies upon accessing a Facility Organization entity's openly available resources through the internet to bring an item to their local computing clusters. Dane, Bryan, Anita, and Palmer each turn to FO entities to obtain climate models. This dynamic with the group's data and model collection efforts emerges because creating climate models from scratch is an undertaking more complex than any single graduate student or even PG would typically have the resources to undertake. It is more time and cost effective to start this group's data journeys by leveraging a product from another entity. Hank noted that among his twenty plus students only one crafted their own model and it took them nine

years to finish a PhD (longer than the typical 5–6 years our interviewees noted). Instead these researchers rely upon the resources of Facility Organization entities such as NASA, the US National Oceanic and Atmospheric Administration (NOAA), or US Department of Energy funded Program for Climate Model Diagnosis and Intercomparison (PCMDI) that support the global infrastructures of climate knowledge. These FO entities sustain their respective climate models or satellite datasets as resources, regularly publishing updates and details about changes widely to the community, while embedding varying practices and assumptions in these complex software assemblages.

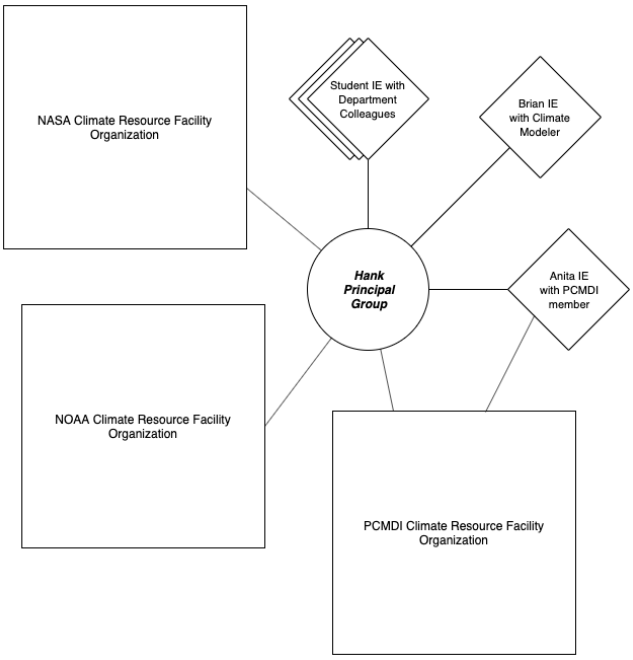

Figure 2. Hank's PG invokes climate resource FOs for data and models while various Intermittent Exchanges are invoked as students conduct work and assess hypotheses.

We also learned that some sources of data that will be integrated into these models come from other researchers conducting field work. Students acquire this data by invoking an Intermittent Exchange through email or other direct communication with the purpose of accessing such a resource and obtaining any needed help. Anita, for example, was writing python scripts to pull data from the beta version of the PCMDI FO's system. During her testing process she enacted an IE with a member of this FO to work through bug testing processes as she probed the FO's data system, even sharing her unfolding code. This fleeting engagement was created to acquire and solidify access to a resource but not sustained over time and did not have a preexisting relationship to draw upon. Her engagement with the FO required synergizing work to develop a relationship necessary for aligning and leveraging resources. Another student, Bryan, also noted how the creator of a model he uses for his work puts in the effort to email known users when an update or bug fix is pushed out. This individual model builder's effort contrasts with the structured efforts of the FO and demonstrates a contrasting ad hoc, fleeting form of organizing to ensure work can consistently be accomplished by interested parties.

> So the guy who wrote this model occasionally will send out emails when someone finds a bug in it. He keeps sort of a list there of the people – that's one of the reasons like we ask for permission to use the model so that he knows who's using it. Then he can send out an e-mail and say like, 'Oh, either I or someone else has found this bug in the model so I've made an update. So if you are using an older version, you'll want to fix this for any future simulations you're doing.' (Bryan, PhD student).

Gathering climate models and datasets melds into the longer-term work of **processing and analyzing data**. Once members of the PG have assembled resources they will iteratively clean data and select variables necessary to answer their research questions—what we have previously described as data processing work (Paine et al., 2015) but extends into analysis work in practice. Through this work these PhD students will generate outputs and discuss them with Hank or another colleague then iteratively adjust the model to test their hypotheses. They will have to unpack the embedded historical journey of the resource before it reached this point, then adapt the model for the new analysis they are carving out. One such key task is examining the different sources for individual data points in a model and assessing their ability to produce quality, relevant data for the question at hand. This data processing effort is necessary to set the stage for scientifically appropriate analyses as Dane explains.

> So a lot of my current project is going through and saying, "Okay, to know something about this particular physical variable, we should really use this data this satellite, because this satellite is a radar [sensor] and therefore is sensitive to ice content," for instance. But we wouldn't want to use to track liquid water because liquid water is not very bright and not very reflective in the radar, so a lot of it is hypothesis driven. (Dane, PhD student).

The work of Hank's PG is collaborative, just among a few people rather than part of a large, amorphous organization of researchers. This first case is relatively simple but illustrates how Intermittent Exchange entities planned permanence (how long the arrangements are planned to endure over time) and nascence (how new the collaboration is and how well routines are or are not established) may be ephemeral. IEs emerge and are invoked strategically when needed, rather than organizing into Sustained Aggregations. The doctoral students—Dane, Bryan, Anita, and Palmer—work on their projects and discuss results with Hank. Anita notes that she shares an office and some computing resources with Hank's other students, but they don't work on the same projects since everyone is studying different climate phenomena using a wide array of climate models and datasets. Overall these climate science researchers form ephemeral Intermittent Exchanges with committee members, other colleagues in their department, and occasionally researchers beyond Seattle. The examples of IEs described by students in Hank's PG are not continually sustained interactively on a day-to-day basis. They solidify at disconnected moments when findings or difficulties need to be discussed, particularly when someone is working out details about a piece of data in a model.

5.2 Waldo: A Marine Geophysics Sustained Aggregation

With Waldo's marine geophysics and underseas seismology PG, we see continual collaboration between his PG and collaborators (another PG) at another university in the Pacific Northwest, Fig. 3. This enduring collaboration is a Sustained Aggregation, the Underseas Seismology SA, that has existed for many years across different grant funded projects, unlike the fleeting IEs we found in Hank's PG, yet still faces nascent challenges as well as ebbs and flows to the relationships and forms of organizing invoked among different members. The Principal Investigators create a stable

base upon which varying arrangements of members come and go over time, demonstrating a case of planned permanence with long-term intentions without a codified organizational structure. During our study Waldo's PG was focused on working with data from an ocean expedition to map the structure of the Earth's crust off of the Pacific Northwest coast. We refer to this as the Ridge Experiment.

Collecting data for the Ridge experiment is a moment offlux for Waldo's PG and the Underseas Seismology SA where the most visible engagement with entities beyond the SA occurs by invoking an Ocean Seismometer Facility Organization. Once data is collected the Ocean Seismometer FO in turn invokes the IRIS FO to deposit data. Data collection work requires organizing varying pieces of complex work using multiple overlapping entities and their processes, resources, and members to temporarily align and stabilize a productive human infrastructure. The invocation of the Ocean Seismometer FO demonstrates a case of significant engagement and interaction with an FO's members and resources as members of the two PGs in the SAwork with FO staff twenty four hours a day during an ocean cruise to craft a dataset that will be long lived. This contrasts with the more ephemeral, simple alignment with the IRIS FO where the collected data is deposited through a computational system once without involving FO staff directly. This latter situation is akin to what we just saw with Hank's PG and their use of climate FOs but depositing a data resource into instead of collecting it from the FO.

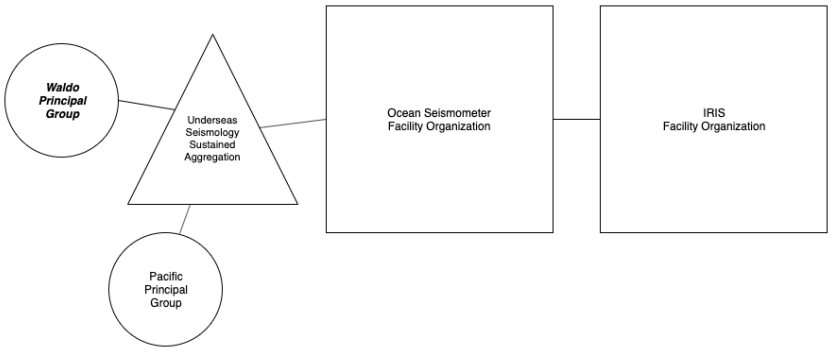

Figure 3. Waldo's PG sustains the Underseas Seismology SA with the Pacific PG. For the Ridge Experiment the SA invokes the Ocean Seismometer FO to collect data. Upon completion of the ocean expedition the Ocean Seismometer FO is mandated by its funding agency to place a copy of the collected data in the IRIS FO as a public resource in addition to providing the SA with a copy.

The Underseas Seismology SA relies on competitive grants received from their US funding agency to temporarily align with the Ocean Seismometer FO entity for a few weeks on a data collection cruise. The funding agency contracts with this FO to operate and maintain a large research ship as well as the seismometers or other research equipment that are brought to sites being studied. The funding agency's proposals are designed with a process where a PI requests the use of a particular set of instruments for a cruise in a specified location where these researchers shoot off a seismic source (air gun) from the ship. If a proposal is funded, then the PI and their collaborators join the vessel and its staff to collect data during a carefully scheduled cruise.

> … we used a seismic source that was on a ship, and we used ocean bottom seismometers that were displayed in a network around the ridge axis. We made explosions over a period of several weeks at sea. And then we are mapping how the sonic waves travel through the earth's crust in the vicinity of the network that we've built. … So I think

> the [Ridge] experiment [data] costs about $5 million dollars to collect. We'll study it for 15 years. (Dahl, PhD student).

This formal organizational process enables Waldo's PG and the Underseas Seismology SA to leverage the FO's resources. We categorize this as instance an example of a Facility Organization and not a Federation because this entity is purely contracted to enable PGs or SAs to collect data. The FO is not helping answer research questions that the Ridge Experiment PIs have developed. This multi-million dollar experiment's cruise results in data that Waldo's PG and their SA colleagues can subsequently spend more than a decade processing and analyzing in various forms.

During the data collection cruise, members of the Sustained Aggregation joined the crew of the FO vessel where FO technicians helped them shoot air cannons into the ocean so that the seismometers can record the reflected sound waves. The technicians employed by the FO maintain the seismometers and are in charge of verifying the data collected at sea. SA members had to balance their research needs with rules for where and when air cannons can be shot to protect wildlife. In the field quality control by members of the PG and FO is essential to ensuring a viable product for long-term research yet the individuals from the FO do not have any involvement in the subsequent work with the data produced, they don't have a long-term stake in the experiment, like they would if this were an example of a Federation entity.

> There are technicians on the ship that are making sure that each shot has the same sound to it, the same source. … And once we start collecting the instruments we have to bring the digital waveforms into a computer and take a quick look to make sure that it recorded everything that we thought it should record. (Rollin, PhD student).

The funding mandates for data collected by the Ridge Experiment's expedition require the raw products be shared, whether members of the SAwant this to happen immediately or not. The FO's seismometer technicians deposit the raw data by invoking the openly accessible Incorporated Research Institutions for Seismology (IRIS) FO data repository. IRIS is a non-profit Facility Organization supported by universities in the United States dedicated to sustaining resources for seismology researchers. This repository is similar to the FO entities providing climate models in Hank's PG as part of a global knowledge infrastructure accessible to a wide ranging community.

Data processing is an undertaking between Waldo's PG and their SA colleagues that can also blur the line into analysis efforts. The initial verification task completed by FO technicians noted above is processing work but necessarily conducted during the cruise so that revised data could be collected if needed. Within the SA data processing work is a variety of activities undertaken by the PhD students, including cleaning different parts of the dataset (e.g., mapping & correcting the locations of seismometers, marking bad data, etc.) and verifying that signals captured are mapped to correct seismometer locations. Subsequently different students will select subsets of data to analyze that is relevant to their given research tasks. Rollin explained how he had to manually process 90,000 air gun shots to "pick" the correct arrival time of the soundwave at the seismometer before he could proceed to analyzing it in a computational model he was developing. Rollin noted how one of the PIs elsewhere in the SAwould engage him in detailed discussions about the particular picks he was making, conveying his experience with this type of data to help Rollin craft the best product possible. Rollin noted how this process has "a learning curve about, well, you shouldn't have made this pick so sharply" and that this PI from the SA might suggest that Rollin "adjust [a pick] in order to make it a little bit smoother." These two individuals had a productive, sustained relationship through the established SA since Rollin had been a member for a few years

by the time he undertook this work. This contrasts with the ephemeral Intermittent Exchanges we saw with students in Hank's PG engaging with individuals outside the PG as they had to process data using FO resources.

Data analysis work involves Waldo's PG and the SA segmenting the processed data to answer the multiple research questions their project is asking. Frequently this will entail re-processing elements as their analyses unfold and discussions take place among various members. The forms and nature of data at any point in time are the result of dynamic interaction, including but not limited to dialogue, between researchers across these entities and the decisions and knowledge they express through software laden actions. Dahl, Rollin, or other members of the SAwho are working on the experiment divvy up the large dataset and iteratively analyze elements of it over time, each taking particular subsets that often rely upon the processing and analysis work other individuals complete, with common research objectives depending on which students join and leave the individual PGs. Rollin's research uses Ridge data to model the Earth's crust from the ocean bottom down for the first few kilometers. Dahl's research is aligned so that he leverages Rollin's upper crust modeling to study the segment of crust immediately below the uppermost down to the mantle. This work is necessary in turn for a member of another PG in the SA to engage in other work with the data.

Waldo's PG works closely with another PG as part of a long-term Sustained Aggregation throughout data collection, processing, and analysis while invoking one FO directly, and another implicitly through the actions of the Ocean Seismometer FO, in the data collection process. Waldo and his PI colleagues have a longstanding relationship to work to acquire funds and produce resources. The PIs who invoked this Sustained Aggregation and undertook the Ridge Experiment leverage their decades long relationships and their continuing desire to have their respective PGs work together to carry on and use this Sustained Aggregation, even as students within come and go as their respective research careers emerge and transform over time. In our effort to categorize this form of organizing it might at first glance be reasonable to perceive this work as a Federation entity. In our inquiry, however, there was not a formalized organizational structure in place nor a charter specifying rules for membership. Waldo and his two PI colleagues are choosing to maintain and sustain alignment of a close set of working relationships between their respective groups.

5.3 Martin: Investigating HIV with Federations, IEs, and FOs

The work of Martin's PG invokes multiple different entities with varying degrees of planned permanence and formality of organizing to collect and analyze data to study the evolution of HIV, Fig. 4. Martin's PG conducts molecular and computational work where individuals work with physical samples (e.g., blood and plasma) to produce genetic sequences that can be analyzed computationally. This microbiological work is conducted as part of different Federation entities over time, but the day-to-day activities unfold within the context of this PG or in alignment with Intermittent Exchanges. The Federations handle work ranging from enrolling HIV infected patients in cohort studies and regularly collecting blood samples to analyzing blood samples for different phenomena and developing new vaccines. The various IEs are created when a resource is needed to accomplish a particular research activity, whether using a costly sequencing machine or using outside statistical expertise for analyses that are ultimately not sustained over time.

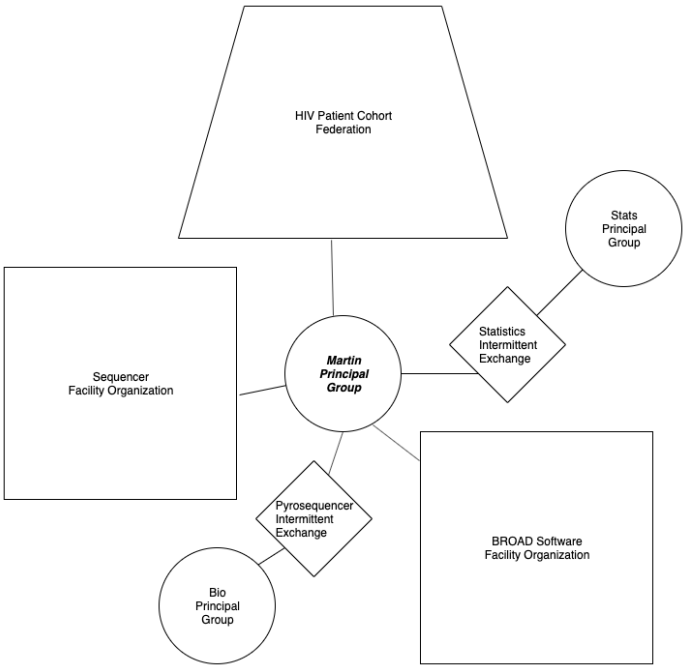

Figure 4. Martin's PG is part of an HIV Patient Cohort Federation which gathers patients and collects blood samples. It has other PGs beyond the scope of our study. During data collection Martin's PG invokes a Sequencer FO and in other situations a Pyrosequencer IE with the collocated Bio PG to access sequencing machines. For data processing Martin's PG either develops software internally or invokes an FO such as the BROAD Software FO. When analyzing data Martin's PG invokes an IE with a Stats PG for certain statistical analyses.

Martin's PG invokes Federation, FO, and IE entities in their work **collecting data**. This process begins with a research scientist in Martin's PG examining records kept by one of the Federation entities they work with to see if a sample necessary to address a particular research question is available and contains enough quantity of virus to make it viable and cost-effective to use. The PG member does this by examining the metadata stored by the Federation. If the research scientist proceeds with a given sample, then they will use one of the different wet lab protocols (Lynch 2002) that Martin's PG has designed to "work it up" so that it can be sequenced, a point in which they invoke additional entities.

> Then what I do is I essentially go look up how likely this sample is going to give us PCR positives, and depending on that, if it's a high-viral load, I just go along with our general protocol. If it's a very, very low viral load, which we think that it's gonna be very template limiting, then I have to take a few alternate kind of testing steps. (Mony, research scientist).

Martin's PG does not directly own expensive DNA sequencing machines which consequently requires the research scientists invoke another entity to accomplish this task. One approach for Martin's PG is to invoke a Facility Organization entity every time through a commercial transaction. This requires the expenditure of grant funds and specification of requirements for the sequencing to be completed, then waiting for this company to finish the work. This may be a variable experience depending on the sequence to be sampled, cost quoted, and how busy a commercial FO entity is at the time. New "pyrosequencing" techniques emerged in the mid-2000s promising increased data volumes opening up new scientific opportunities and Martin's PG did

not purchase one of these expensive machines or rely upon a commercial sequencing company. Instead Martin's PG leveraged an existing relationship with a colleague, Professor A with the Bio PG who did purchase one of these expensive machines, by invoking the Pyrosequencer IE. A research scientist in Martin's PG was trained to run the machine leveraged through the IE to produce data necessary in a temporary alignment.

Martin's PG undertakes **data processing** work within the PG but uses software from a Facility Organization when it suits the PG's research needs. Members of Martin's PG will have to spend significant amounts of time working to fix insertion and deletion errors that arise because of flaws in this pyrosequencing technology and aligning the overall sequence using automated software and often painful manual intervention. This data processing work entails both creating new software pipelines and adopting software from the larger microbiology community. For their first pyrosequencing project Sharvani, one of Martin's PhD students, developed a new software pipeline. For their second pyrosequencing project the PG tested a variety of pipeline software before selecting one that is publicly available on the internet from a Facility Organization entity, the Broad Institute in Massachusetts. In this situation Martin's PG chose to use the Broad FO's freely available software to support their local work rather than adapt Sharvani's pipeline or write something from scratch due to the time saved by leveraging this external resource.

> So the cleaning process during this one is actually a sort of commercial one. It was made by the Broad Institute. They have a set of sort of scripts and software that you can download that have been validated that will go through and clean your data, and basically what it gives you at the very end is things that it has determined that are real sequences. (Elisa, research scientist).

Data analysis work results in Martin's PG engaging with the larger Federation entities their projects are a part of as well as different Intermittent Exchanges. Processed and cleaned pyrosequenced data can be analyzed for particular questions, such as comparing sequences of vaccine recipients versus those of a placebo group to see how the virus was forced to evolve in the vaccinated patients. Members of Martin's PG perform common analysis tasks using a variety of software scripts and Excel spreadsheets. Such tasks include calculating basic evolutionary details about each HIV sequence, the quantity of virus in the sample, and an assessment of how well a patient's immune system was fighting the virus. Across different projects members of the PG will divide up processing and analysis tasks in parallel to iteratively work through the large quantities of data produced. The PG as a whole regularly shares in-progress work through multiple different weekly meetings, including a computational analysis group meeting with roundtable sessions where emerging results and challenges were discussed.

Analyzing pyrosequenced data often requires crucial immunological metadata about a patient sample that requires Martin's PG engage with other PGs in their Federations to acquire since they do not conduct immunology work. Sharvani noted there is a "bureaucratic process" she would have to go through where she had to "ask permission formally from them to give me immunological data, and they have to, like, have [Martin] sign off on it saying that this person is valid, this is, you know, part of [Federation] work." Here Sharvani and Martin leverage membership in this Federation by invoking a defined process to acquire siloed data needed to answer questions and attest to its use for Federation sanctioned research problems.

Conducting other analysis tasks frequently requires the invocation of an Intermittent Exchange with statisticians at another Seattle research center. After Sharvani, Elisa, or another computational

researcher in the laboratory finishes processing a project's sequence data they will share these files with their statistician collaborators. Through this IE different aspects to research questions will be examined and discussed in the process of writing up results. Members of Martin's PG leverage relationships with other local researchers to align research interests and resources (statistical knowledge, cleaned data) so that together a hypothesis can be investigated through this Intermittent Exchange entity.

Throughout this case we see Intermittent Exchanges, Federations, and Facility Organizations invoked with relationships engaged and sustained ranging from scientific colleagues to commercial companies. Similar to Hank's PG we find that Martin's PG uses Facility Organization resources for data collection as well as data processing, in addition to an IE with another local biology PG. Our data did not suggest a Sustained Aggregation since the organizing purpose at hand was oriented around access to the pyrosequencer without any demonstrated intellectual engagement at the time. This is our only case where an FO is invoked for an activity after data collection work and only to acquire a piece of publicly available software. Our inquiry also did not directly capture much about the dynamics of any of the HIV cohort Federations that this PG works with since we were unable to attend any Federation meetings or engage with other PGs in this entity. This is in contrast to our fourth and final case of Magnus whose PG works with multiple nested Federations on an intimate, day-to-day basis perpetually.

5.4 Magnus: Cosmology with Nested Coordinative Entities

Our final case is that of Magnus's PG and their cosmology research which portrays our only example where multiple forms of coordinative entities were created and employed in a nested or overlapping manner with each having long-term planned permanence. Magnus's PG works by invoking the Widefield Radio Telescope (WRT) Federation, that has four distinct sub-Federations (A, B, C, and EoR), and multiple Sustained Aggregations in a multinational undertaking building a radio telescope and high-precision data analysis software, Fig. 5. We have previously referred to this overarching Federation as the WRT project (Paine 2016; Paine and Lee 2014, 2017; Paine et al., 2015). Magnus's PG, the SAs, and EoR Federation all contribute to and rely upon the WRT Federation's telescope built on land maintained by the Southern Hemisphere Radio Observatory FO. The Southern Hemisphere Radio Observatory FO is invoked to manage land, electricity, fiber optic connections, and to address political concerns related to aboriginal ownership of this remote desert location. Through these forms of organizing these cosmologists produce petabytes of data as the raw material for their complex software pipelines and analyses.

Following the threads of Magnus PG's cosmology work we see constant engagement with the US EoR and International EoR SAs organized by members of the Epoch of Reionization (EoR) sub-Federation. Both of these SAs were organized with members who are from the same geographical area. Our inquiry initially focused on Magnus's PG but extended outward into the US EoR SA and EoR Federation. The planned permanence of all of these entities is long-term, even as the arrangements of individuals we studied undulated as careers and research agendas shifted over time. The many individual PGs throughout these entities have to continuously work to maintain and sustain alignment of their different goals and tasks among and through this web of coordinative entities to collectively accomplish this complex, large scientific undertaking.

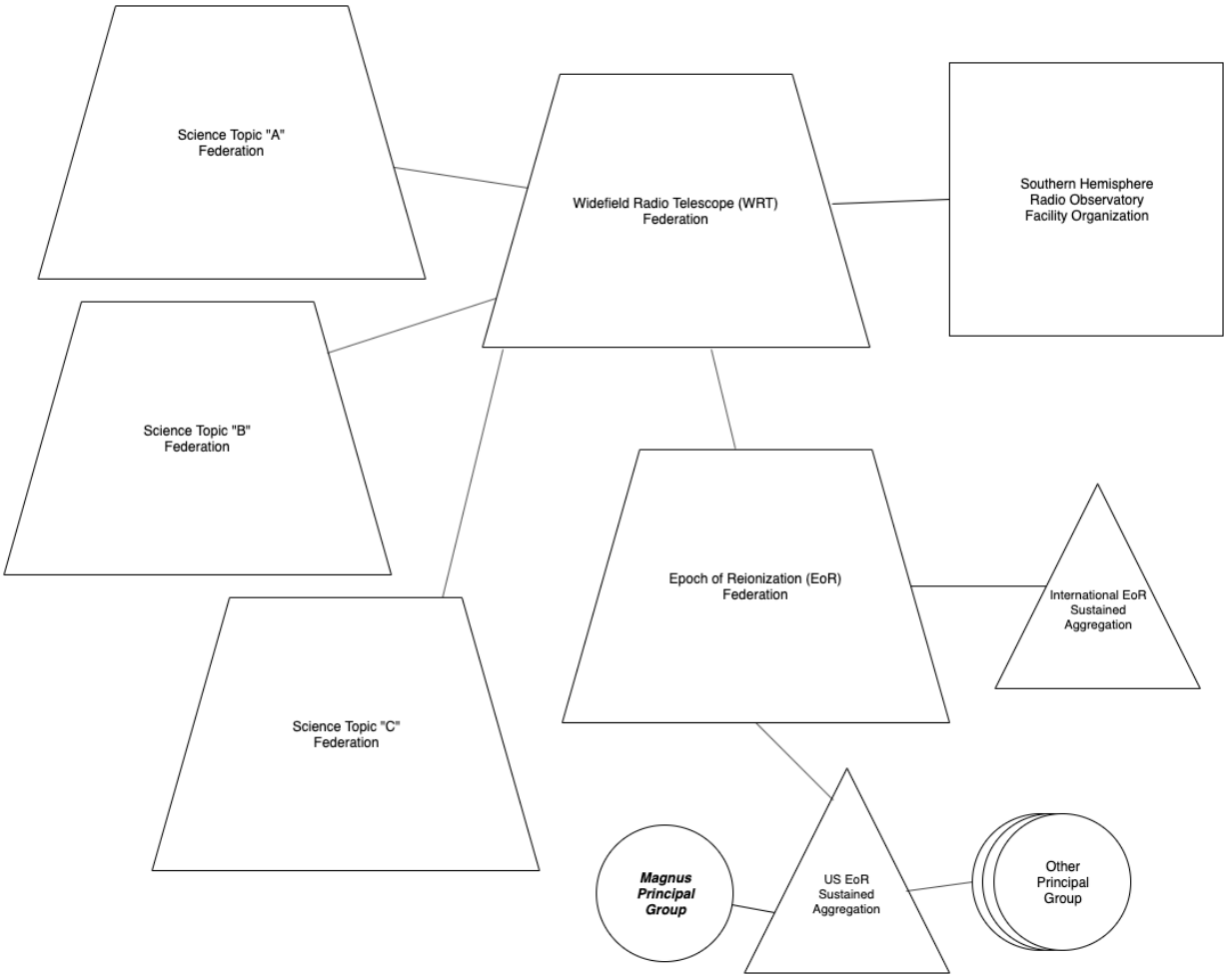

Figure 5. Overview of the Widefield Radio Telescope Federation with different sub-Federations and their constituent PGs and Sustained Aggregations. The WRT Federation invokes the Southern Hemisphere Radio Observatory FO to obtain a site for its physical telescope along with resources like electricity and fiber optic connections to sustain the instrument. We did not study the Science Topic A, B, or C Federations.

We see nesting and interconnected, embedded relationships among these entities firsthand, finding Magnus's PG iteratively exploring subsets of data by comparing results with colleagues from the US EoR SA, International EoR SA, and EoR Federation entities and to refine the software instrument at the core of their work. The members of these entities are all pursuing Epoch of Reionization science and developing rules, practices, and resources for EoR data access and use in the EoR Federation (that any scientist would be subject to, regardless of entity affiliation) by extending the WRT Federation's general policies around its resources. The SAs within the EoR Federation can also end up developing and sustaining distinct resources for their more narrowly bounded entity. One of the founding PGs in the US EoR SA was tasked with procuring a computing cluster. PG members from across the EoR SA's leveraged this system as the computational site for data processing and analysis (much of their software work was shown to us while remotely logged into this cluster) and helped to sustain it by maintaining the system's software configuration over time. Even with the general stability of these organizational arrangements during the period of our

study the particular PGs in the US EoR SA, and the actively contributing individuals within them, changed as some PI's research agendas shifted with new projects or funding streams. For example, the PG that procured the computing cluster in the US EoR SA pivoted to another telescope project after a few years and took this shared resource with them.

Data collection finds Magnus's PG helping to craft processes and systems used by the overall EoR Federation, which relies upon the Southern Hemisphere Radio Observatory FO, to operate the telescope and move the resulting data products to computing systems around the world, although most of this work was an upfront task and intentionally not sustained once running smoothly. This PG contributed in part when Brianna was volunteered by Magnus to develop pieces of the whole telescope's monitoring and control software, contributing service labor to the EoR Federation as well as the overall WRT Federation to enable reliable data collection. To collect EoR data observing time is allocated to the EoR Federation in accordance with the WRT Federation's policies. A member of one of the PGs will then program the telescope's control system to automatically capture and archive data on WRT Federation systems based on the configuration. Originally the EoR Federation developed a distributed practice where a member of the US EoR SAwould monitor observations in real time looking for any spurious details with the instrument. The following day a researcher in the International EoR SA would do a quick validation of the observing night's collected data. This practice broke down and dissolved after a short period in part because members of the EoR Federation were not fully adhering to the specified practice but also because they deemed the instrument stable enough that it was not necessary to fully monitor each observing night. Through subsequent processing and analysis these scientists could simply determine when to throw out part or all of an observing night's data. This was less labor intensive than actively monitoring unfolding data collection.

> … now that it's an operating instrument, there is somebody on the ops team who's charged with laying out the schedule about who gets to observe when. And typically, that's on a per-night basis, we don't usually have more than one group trying to operate in the same night. … And then the person who's in charge of EoR observing … schedules them then. And then everybody in the EoR [Federation] has responsibility to monitor the instrument and make sure things are going right. (Brianna, post-doctoral researcher).

Over time members of Magnus's PG and the US EoR SA would end up directly copying the EoR data products produced from the WRT Federation data archive to computing systems they directly work on (the US EoR SA's cluster or local laptops and desktops) by writing and maintaining different software scripts.

Processing and analyzing data is an effort that Magnus's PG undertakes with other entities in the EoR Federation using two high-precision data-analysis software pipelines that are able to exchange myriad intermediate data products. Magnus's PG are the developers of one of the two software pipelines (the other developed by a few PGs in the International EoR SA), that effectively become software telescopes (Paine 2016). Igor and Brianna, both post-doctoral researchers, were responsible for implementing two primary components of the PG's software over a multi-year period. Abner, Peg, and Nima as doctoral students in the PG were tasked with executing elements of this software pipeline to produce products for examination and eventual refinement of the code. Each pipeline employs intentionally different scientific methods to enable Magnus's PG and their EoR Federation colleagues to debate and assess distinct approaches to this science. In the simplest

form, executing these software pipelines with EoR data is the work to process it, while assessing the outputs is analysis.

The goal of both Magnus's PG and the EoR Federation is to produce a statistical power spectrum measurement through their processing and analysis work. These entities do so by evaluating whether a change to their software pipelines improves power spectrum outputs conveyed through various plots. Using a standardized "golden dataset" Magnus's PG, the US EoR SA, and International EoR SA each run their pipelines, exchanging various intermediate products to isolate effects that emerge from various methodological choices embedded in each set of software. As they refine their pipeline code Magnus's PG will first process the golden dataset and compare new plots to a past understood revision. For complex issues they then have discussions during weekly videoconferences with US EoR SA colleagues and eventually share more widely with EoR Federation colleagues. This increasing scope of engagement is exemplified by Magnus PG's work to understand and handle the fourth line bug unpacked in Paine and Lee (2017).

> And a lot of the tests that we've done along the way has been more of testing out our analysis on that set [golden] of data. So sometimes if we change something in the code somewhere, we want to see how does it affect the power spectrum, so you'd run it on that standard set of data to compare what the output was. We see whether it improved or hurt the power spectrum. (Abner, PhD student).

Intimately working as part of nested SAs and Federations distributed across the world enables Magnus's PG to create and improve a nascent software telescope and further their local research goals. Magnus's PG continuously draws upon relationships across these forms of organizing. Each particular entity we have identified (Magnus's PG, the US EoR and International SAs, EoR Federation, WRT Federation) endured throughout the period of our study. Peering inside each instantiated entity as a form of organizing however we saw individual members change when students graduated, PIs left for different telescope projects, or funding for a given PG's participation ran out. Magnus's PG perpetually had to work to sustain alignment and grow relationships with these arrangements of globally dispersed scientists by working as part of the WRT Federation. This type of work is a foundational organizing principle of this evolving social world, illustrating a "big science" endeavor where multiple entities are invoked to affect and shape work, and Magnus's PG continually plays a significant role. By following Magnus's PG we surfaced how one dynamic PG can invoke and sustain relationships among many entities on a day-to-day basis to be able to get work done. Magnus's PG was instrumental in the emergence of the particular entities we identified from the time they were created (Magnus being a PI who helped found the WRT Federation) onward through their temporary solidifications, maintenance, and perpetual change (contributing through expansions of the telescope and changes to the WRT Federation's organizational structure). The relationships formed and invoked vary depending on the task at hand, sometimes focused on the creation of a usable dataset and others on iterating one particular element of a complex software telescope, but these relationships are readily able to be aligned and leveraged thanks to the creation and sustainment of these many overlapping or nested coordinative entities.

6 Discussion

Coordinative entities are a mechanism to decompose and characterize emergent ways that PIs and their groups organize their work. These entities enable us to see how different coordinated actions are connected, or how coordinated actions are cobbled together to form a larger, more complex, composite coordinated action. Other actors can and do also work with and through these entities, but we have focused here on PIs as their roles require institutional entrepreneurship. Employing our typology of coordinative entities refocuses our understanding and allows us to step back and draw out similar and diverging arrangements that facilitate data intensive work in the complex landscape of scientific groups. In this paper we have not attempted to produce a comprehensive list but rather present a first attempt to identify some repeating types that can be useful to the analysis of coordinated work and innovation. If we take seriously the idea that inspecting the sociality around coordinative actions (such as the entities described in this paper) is important for disentangling scientific work, then understanding how these entities interact will provide more ways to understand and scope design projects and spaces and will help us grasp the landscape of stakeholders and the relationships between them.

Scholarship has highlighted how scientist's forms of organizing expanded with the adoption of distributed, internet enabled tools from local groups to temporally, geographically, and intellectually diffuse collaborations such as team science (National Research Council 2015) and citizen science (Bonney et al., 2009; Wiggins 2013; Wiggins and Crowston 2010) where stakeholders can often have conflicting interests (J. Olson et al., 2008b; Ribes and Finholt 2009; Velden 2013). While the scientists in our study organized themselves in very different ways during data collection, processing, and analysis to meet their immediate and longer term needs, our research found some common recurring entity forms. The collaborative science of these researchers is an intricate affair where Principal Groups will participate in varying kinds of research projects, many of which require participating in different collaborations at the same, different, or overlapping times. Over the course of doing research, different entities were created and then connected together (invoked). Our examples illustrate a continuum from close-knit and local collaboration at one end to a loose and dispersed collaboration at another. To undertake data intensive science these researchers are constantly creating and reshaping different common fields of work to answer problems, just as architects construct fields of work through the unfolding processes of design, planning, and construction (Schmidt and Wagner 2004).

Our goal with this work is not to create a reductionist model of scientific collaboration, but rather to begin to produce a conceptual framework that includes meso-level phenomena and enables us to see and scope design spaces that span individuals, groups, organizations, and infrastructures. Science and Technology Studies researchers Wyatt and Balmer (2007), citing Newstead et al. (2003), have noted that many cultural geographers who work daily with matters of scale have "largely abandoned the notion that scales such as local, national, continental, or global are fixed and that different actors invoke different scales to make sense of their actions." While this may at first seem to be a critique of the development and discussion of meso-level or middle range theories, the argument falls flat if we consider that "meso-level" could not only refer to multiple overlapping levels and that all "levels" could potentially overlap. In other words we can reject the micro-macro binary and rather than positing a micro-meso-macro tertiary or

flattening everything as does Latour (1987, 2005), we instead consider the meso as another important and complementary lens for better differentiating particular phenomenon. While these coordinative entities are useful forms of organizing, the real power of the actors we follow are their ability to undertake different forms of synergizing to create new linkages—and effectively new organizations—that vary greatly in terms of their formality, tasks, purpose, and planned purpose. In fact these linkages are made very visible by focusing on meso-level phenomenon and furthermore point to how connections are made to institutional actors and to better enable scholars to show how individuals could act as "institutional entrepreneurs" and how, when, and why actors engage other actors to make new things happen (DiMaggio 1988).

Organizational sociologists Fligstein and McAdam (2012) note that meso-level social orders, what they term strategic action fields, are the "basic structural building block of modern political/organization life in the economy, civil society, and the state." Fligstein and McAdam assert that established theories, such as the foundational work of Giddens and Bourdieu, while useful, are also very vague about many aspects of the actual dynamics of meso-level action where actors work. Fligstein and McAdam critique and build upon Giddens saying that in order to understand how actor's actions not only perpetuate and stabilize but also change how things are done it is critical to understand actor's particular stakes and their moves and motivations to control strategic action fields. Addressing Bourdieu (1984) and Bourdieu and Wacquant (1992), Fligstein and McAdam note that:

> Actors in Bourdieu's theory are generally only responsible to themselves and motivated by a desire to advance their interests within the constraints of the situations in which they find themselves. But fields also turn more centrally on coordinated action, which requires actors not to simply focus on their position in a field but to seek cooperation with others by taking the role of the other and framing lines of action that appeal to others in the field. We view these collective dynamics as complementary to the generally individual action that is Bourdieu's central concern. (Fligstein and McAdam 2012, p. 25).

Relatedly, Gergen's (2010) perspectives on social constructionist theory, and its applications to practices of social change, in similar fashion emphasizes that in studying organizational practice researchers should be redirecting attention away from the traits of individual people, technologies, or artifacts and towards the relationships between entities. Given the focus in CSCW to inform design, the need to be able to understand, name, and clarify entities and the dynamics of individual and entity actors has important practical implications. Theories that open space for close inspection of middle range actors and actions such as coordinative entities help us undertake such a task by offering a different way of disambiguating relationships among researchers and organizations.

6.1 Theories of the Middle in CSCW

The field of CSCW, as with related fields like organizational sociology, has made efforts to understand connection strategies of actors and social stakes of organizing. In CSCWAckerman et al. (2008) in particular advocate for development of theories of the middle, "small-scale theories that would allow CSCW and adjacent fields to move forward in a more systematic and less hit-or-miss way." Our prior work developing the model of coordinated action is one such endeavor (Lee and Paine 2015). With this work we articulated a model of CSCW scholarship which encompasses not only narrowly bounded goal-direct coordinated actions but also diffuse, messy engagements

where people work together but not necessarily with shared goals. Coordinative entities complement this as a meso-level theory enabling us to see how different arrangements of particular types of coordinated actions are invoked, sustained, and dissolved in varying types of scientific research and its activities, whether in the work of producing data by leveraging the resources of multiple different entities or the processing and analysis of this product in more bounded situations.

The importance of the concept of synergizing (Bietz et al., 2010) is that it describes the creation of a *common field of work* (Schmidt and Simone 1996) on which an ensemble can enact changes. Synergizing was theorized primarily to address how individuals can act to form interorganizational relationships. A local, temporary alignment of practices (Yasuoka 2009, 2015) must be strengthened so as to be sustained as a thing longer term. We see the importance of the development of common fields of work among coordinative entities through the various examples of local, and often temporary, alignment of practices in our cases. When Martin's PG leverages a relationship with the Bio PG they are temporarily aligning their interests to access a vital instrument resource for their work. This was also the case with Anita's invocation of an Intermittent Exchange with the PCMDI Facility Organization when trying to obtain climate models.

In time, the work accomplished by invoking entities shifts from the initial creation of common fields of work towards some form of articulation work (Strauss 1988) as coordinated actions are sustained to advance shared research agendas. Coordinative entities are in practice the articulation of articulation work. Magnus's PG sustains engagement on a daily basis with multiple nested coordinative entities and in doing so requires various forms of articulation work that we can follow, from arranging regular meetings around particular types of tasks to managing the movement of data and software products among computing systems. Gerson elaborated upon Strauss's formulation of articulation work and offered the complementary notions of local articulation, making sure resources are in place and functioning when and where needed locally, and metawork, putting together task clusters and sequences, not necessarily locally (Gerson 2008). Coordinative entities are both a contributor to and a product of these types of articulation and synergizing work. In the course of doing research a Principal Group's members will bring together tasks and lines of work for many research activities. Each research activity in our study entails the formation of new composite organizations, whether during data collection, processing, or analysis.

The scientists in our study conducted different types of research with different constraints, technologies, and institutional contexts. Our four cases span a spectrum of disciplines and engage in diverse modes of investigation (e.g., simulation, observation, and experimental). At the same time these scientists do have some similar concerns and constraints as they are all participating in data-intensive scientific research. The conceptualization of coordinative entities supports our examination of each research activity as an organizational endeavor and supports our examination of PIs and members of their research groups actively forming organizations through synergizing, articulation work, and metawork. Using this typology enables us to see work that is all too frequently understudied, and sometimes barely visible, as comprehensive, multi-sited and/or a phenomena scoped in multiple ways in scientific research—for both the scientists studied and the CSCW researcher.

6.2 Using Coordinative Entities when Studying Data Intensive Science

Our typology of coordinative emerges from our analytical efforts to compare quite different forms of scientific work. Following the work of different scientists, rather than infrastructuring undertakings, we initially had difficulty making the numerous meshes of people, activities, and resources tractable for comparison. Scholarship previously shifted our attention in infrastructuring design work from short term to long term (Karasti et al., 2010; Ribes and Finholt 2009), and now shifting our investigations of scientific collaboration away from singular infrastructuring endeavors is an opportunity to better investigate, understand, and account for the varied forms of organizing and sociality underlying different coordinated actions today. Identifying and tracing how our four cases invoke different coordinative entities across three types of research activities (Table 2) enables us to compare and contrast work conducted in disparate disciplines and begin to better surface the arrangements that help diverse coordinated actions be productive (Bietz et al., 2010)—and potentially reverse synergy where trying to improve productivity leads to failure and dissolution (Langhoff et al., 2018).

The ways different Principal Groups conducted work by engaging with different coordinative entities that overlap, and at times nest, surfaces a complex web of scientific collaboration. Lee et al. (2006) emphasized the multimorphous nature of human infrastructures in scientific work and coordinative entities enrich our understanding of this heterogenous aspect of data intensive work. Fujimura (1996) describes how scientists co-construct their work through "shaping and adjusting materials, instruments, problems, theories and other representations, and social worlds as well as themselves and their laboratories" (p.207). Following the arrangements of coordinative entities in these sites we showed how different ways of organizing and re-organizing over the course of these researcher's constant coconstructive efforts helps them align and leverage varying relationships among people and resources, whether it is for a singular project that we might characterize as an infrastructuring endeavor or for a more fluid, less bounded undertaking.

The representations of the webs that emerge in each of our four site's work (Figs. 2–5) results in abstractions that can help to highlight diverse forms of organizing and would especially do so when generated for multiple points in time and place. With these figures and Table 2 as a baseline we can break down types of data work and lay out the journeys (Bates et al., 2016; Leonelli 2016) or careers (Harper 1997) that these artifacts or objects take among different organizational arrangements, highlighting paths taken and not taken as well as the relationships necessary to this work. Drawing our attention to each type of coordinative entity across data collection, processing, and analysis we see how different PG's work ebbs and flows as fluctuating relationships are invoked and relied upon. Following these arrangements we can disentangle work with data while considering how our own categorization efforts shapes the ways we see and understand the complex co-constructive work of modern data intensive scientific research.

Table 2. Coordinative entities invoked in each case during data collection, processing, and analysis. The dashed line between Processing and Analysis represents the fuzzy boundary between these activities in our data.

	Collection					*Processing*					*Analysis*				
	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
Climate Modeling	X	X			X	X	X				X	X			
Marine Geophysics	X		X		X	X		X			X		X		
HIV Microbiology	X	X		X	X	X				X	X	X		X	
Cosmology	X		X	X	X	X		X	X		X		X	X	

Key

#	*Entity*	*Diagram Shape*
E1	Principal Group	
E2	Intermittent Exchange	
E3	Sustained Aggregation	
E4	Federation	
E5	Facility Organization	

6.3 A View across the Four Cases

By using the five different entity concepts as a lens to see how organizational forms are invoked by members of these different research groups we can begin to make some simple comparisons that can inspire more research questions and studies. As one would expect from the construction of this lens, the Principal Group plays a key role for all the sites represented in all stages of their research projects. We also see that all four groups conducting data-intensive science rely on Facility Organizations for the collection of the data but subsequently only one group works with a Facility Organization for data processing or analysis work. We do not have data on data dissemination, however, so it possible and likely that Facility Organizations come back in to play during that activity given the intention of such activities is to make products widely available over long timespans as research priorities and understandings change.

Cosmology work in Magnus's PG has no identified Intermittent Exchanges, our site that has the most formalized organizational forms with nested Federations. At the opposite end of this spectrum, climate modeling in Hank's PG, which invokes neither Federations nor Sustained Aggregations, has the most reliance on Intermittent Exchanges. Our two sites that invoke Sustained Aggregations do not just invoke those SAs at one particular part of the research process but rather throughout the entire process. It is possible that Sustained Aggregations are not merely long term but also indicative of a more intensive and interwoven way of collaborating than that implied by the other types of coordinative entities. Seeing just two sites invoke Federations (the Magnus cosmology and Martin HIV research PGs) we noted that both cases are in the business of sustained, continuous data production for their fields. This contrasts to the marine geophysics work of Waldo's PG, for example, who produce data during individual cruise events but not continuously over time.

Future explorations of more PGs connected to Magnus's PG might reveal IEs were we able to investigate more ad hoc work that does not fit neatly into any one of the topic-based sub-Federations of the WRT Federations. The organization of work in Hank's PG was noted by members as not the norm among other climate research groups. This PG-oriented focus is how Hank as a PI has organized and sustained his PG's work and this is possible in part because of the complex research products this community's knowledge infrastructure produces and sustains for individual PGs to be able to wield as desired.

These high-level comparisons are rudimentary and preliminary. The focus of this paper is simply to lay out the framework of coordinative entities. Future ethnographic studies, however, could benefit from linking rich description with an analysis using coordinative entities to further disambiguate the diverse qualities, sites, and dynamics of scientific collaboration. Continued work in this direction will also contribute towards deeper and more extensive understanding of how assemblages of scientific collaborations function and change over time to meet immediate data-related needs. This knowledge would be immensely beneficial to CSCW and science policymakers who wish to support the conduct of science. Coordinative entities furthermore enables us to examine patterns in how different types of work are conducted (e.g. in data collection or processing) across ethnographic case studies.

6.3.1 Disentangling Entities Invoked in Data Collection Work

Inspecting the relationships necessary in our four site's efforts to collect data consistently materializes the widest array of coordinative entities in the work we followed (Table 2). How many coordinative entities visibly emerge differs, but each PG relies upon a Facility Organization coordinative entity for resources. The two cases that work with Federations (the Martin and Magnus PGs) invoke at least three entities beyond the PG to collect data while the Hank and Waldo PGs each invoke two other coordinative entities. At first glance it might be reasonable to categorize the scientific endeavors using the dichotomy that emerges with the big or little science trope (Galison and Hevly 1992). Darch and Sands (2015) reframe this dichotomy by asserting work from both scales dynamically affects the others. With our typology of coordinative entities we disentangle this issue from an organizational point of view to begin to bound and categorize the relationships invoked throughout the work of data collection rather than reducing certain activities to big or little scales alone. Seeing the diversity in number of entities helps us explore the ways relationships grow, solidify, and potentially whither in the work to collect (or process, analyze, etc.) data, software, or other emergent infrastructural components.

Taking the case of Hank's PG we see a coordinated action where individual members tackle related but distinct research problems under the auspices of Hank's overarching research agenda. Hank's PG collects data and computational models from Facility Organization entities since the group is primarily a modeling group that uses and modifies publicly available resources to do their research. These FO resources are the products of "big" science endeavors, requiring a vast political machine to generate this key material (Edwards 2010). This is our only case where the PG is fully reliant upon distributed entities to produce the starting materials necessary for their work, contrasting with all of our other cases, even as Hank's PG does have to invoke IEs and FOs to do their work such as Anita's efforts to overcome computational friction (Edwards 2010) when pulling a FO's model into her unfolding work. This work we have categorized as data collection

is in practice a type of data reuse (Faniel and Jacobsen 2010; Rolland and Lee 2013) and the work of Hank's PG is to assess how the products of other entities align to their research questions and at times invoke IEs to successfully complete this task. Collecting and reusing existing resources is possible because the data cultures or economies (Vertesi and Dourish 2011) in the global knowledge infrastructure of climate science result in Facility Organizations chartered to provide such resources.

Our other three cases in contrast are directly involved in the production of observational or experimental data using different instruments. The PGs of Waldo and Martin align themselves with other entities to be able to leverage expensive instruments that they would not otherwise have available. We see that the ways each PG accomplished this differ, where Martin's PG only had to form an Intermittent Exchange with a PG in the same building while Waldo's PG working with Sustained Aggregation colleagues had to undertake processes to align their entities with a Facility Organization and its rules and practices regarding instrument resources. Magnus's PG is our only case intimately involved in the design and development of new instruments (both hardware and software) then using them to collect data in conjunction with multiple overlapping, nested coordinative entities—a highly multimorphous human infrastructure.

6.3.2 Making Invisible Data Processing Work Visible

Data processing work is the often laborious task to transform resources into an analyzable state that as a process can be rife with easily lost changes that shape knowledge being constructed (Paine and Ramakrishnan 2019; Paine et al., 2015; Plantin 2019). Efforts to process and clean data are often work that melds into the background, invisible to outside observers looking at the shiny elements in ecologies of work who are not always paying attention to all of the indicators (Star and Strauss 1999). Following the work of our four cases and the entities each of these PGs invokes to accomplish different tasks we are able to foreground and focus on the complexities of data processing work as an integral scientific activity (Paine et al., 2015). This was our experience as we worked to categorize and disentangle the practices and activities we were seeing emerge in our data.

Data processing work is quite visible work when we focus directly on each PG. The organizational arrangements discernible when examining data processing work across these four sites indicates Waldo and Magnus's PGs engaging with Sustained Aggregations continuously while Hank and Martin's PGs have much more fleeting engagements through Intermittent Exchanges. Each of our case's data processing work is incredibly intricate, bounding up such work with the different analyses to be conducted and relying upon the relationships of these entities. Much of this work is that of cleaning datasets and assembling different datasets into one analyzable form using a variety of pieces of software.

In both the Waldo and Magnus cases, processing data brings about continual discussions about the state of the artifacts being produced and the knowledge different individuals among the entities have at a given time. Rollin as a PhD student in Waldo's PG received feedback and support in their effort to clean a massive seismology dataset through the relationship buoyed by the Underseas Seismology Sustained Aggregation. With the work of Magnus's PG their efforts processing and analyzing data through their software telescope surfaced frequent invocation of the US EoR Sustained Aggregation as this arrangement of researchers in Seattle worked to take petabytes of

telescope observations and craft useful plots. Processing leveraged a US EoR SA PG's computing cluster for computation and the insights of other researchers working as part of this coordinated action.

6.4 Challenges Categorizing Organizational Arrangements

Our effort to identify and categorize common forms of organizing across our field sites is a step towards having more systematic ways of comparing and contrasting different kinds of data intensive science. The typology of coordinative entities enables us to materialize the ways coordinated actions accomplish work. How we categorize or construct these elements is a key decision that requires reflexive conversation with the conceptualizations scientists themselves have about their ways of working and the assortment of coordinative entities we identify. Larkin (2013) reminds us that considering an 'infrastructure' is a categorical act and similarly pinpointing forms of organizing in heterogeneous work is also an attempt at usefully classifying relationships among scientists, resources, and so on.

Karasti and Blomberg (2018) remind us that infrastructure is always relational, emerging and accreting in different ways for different people, while from an outsider's perspective only ever able to appear in fragments, emerging in partial forms depending on how the ethnographer follows connections and discontinuities and bounds the phenomenon and field of inquiry. Employing coordinative entities to frame variable scientific work raises these issues, even when we are not investigating singular infrastructuring undertakings. With Magnus's PG the perpetual nested, overlapping webs of coordinative entities visible to us as outsiders raise questions of where it may be best to begin and/or situate a given analysis. Our inquiry began with this local group in Seattle but in practice over time followed threads around the United States and world, investigating the relationships and contributions of a varied web of individuals helping to advance Epoch of Reionization science.

It is worthwhile to ask how our insights would differ if we had elected to follow the WRT Federation in the vein of a more typical study of an infrastructuring project. Following this larger, multimorphous Federation coordinative entity could have enabled us to focus on the kernel of this research infrastructure (Ribes 2014) or explore the inter-relations of a different web of Principal Groups. Elements of the WRT Federation could have appeared as a Facility Organization entity if the focus of our study was a PG operating like Hank's climate science PG, rather than Magnus's PG which was an integral contributor to the creation and development of this Federation and many of its resources. Ribes (2017) describes two "cohort studies" in HIV/AIDS research that would be categorized as Federations for him as a researcher using our characterization, similar to our case with Martin's PG and their work, yet the data they produce and sustain could be available in a Facility Organization situation to a PG not involved in the original cohort studies. The CAMERA cyberinfrastructure project in Bietz et al. (2010) was a Federation working to build resources, but in a different investigation it may have been a Facility Organization if a group being studied is simply relying upon some resource being made available by the CAMERA project. Further exploring the overlapping space between Federation and Facility Organization entities is an opportunity to clarify the impacts of shifts in perspective by the researcher, as well as insights about the impacts of disparate funding structures (Kaltenbrunner 2017; Kee and Browning 2010).

Overall the membership of Magnus's PG in the WRT Federation, its nested EoR Federation, and the US EoR Sustained Aggregation emerged in our analysis as so bound up with this local group's work that it is questionable whether we could usefully categorize this PG's work all that differently. Had we started our inquiry with a different PG that engages with a larger variety of Federations or SAs not oriented around a focal endeavor then our view would be different and would of course surface other perspectives on this type of work. Further still, we recognize that not every member of a particular PG will be engaging with a Sustained Aggregation on a daily basis, yet in this cosmology case even the PhD student most disconnected from the core software pipeline work ends up relying upon this instrument and its products. Over the course of our dialogue about Magnus PG's work, and the other three cases, we have found that the coordinated actions these scientists work with and through solidify through the relationships they craft among people, resources, and ideas to arrive at doable problems (Fujimura 1996).

Similar challenges can be found in our other cases and with the ways we design inquiries of data intensive science or other amorphous, computationally laden work. Coordinative entities are an initial attempt at helping us be more precise in identifying what is within the scope of studies we as investigators of data intensive science are undertaking.

7 Conclusion

This articulation of types of coordinative entities is only one step toward decomposing and disambiguating scientific collaboration. We find these coordinative entities, however, are a necessary step to be able to compare the work of these different scientists with their diverse cultures, practices, and methods as they enact different locally rooted infrastructural components among varying social worlds. Prior examinations of human infrastructures and synergizing work begin to draw our attention to the varying relationships scientists invoke in their complex research creating common fields of work (Bietz et al., 2010; Lee et al., 2006). Our findings here advance their utility by helping CSCW scholars characterize some of the varying organizational arrangements data intensive scientists employ across projects over time. Rather than study just particular infrastructuring projects, we focused on the forms of organizing crafted through relationships among individuals, resources, and entities *across project and activity boundaries* to unpack how data intensive science is accomplished. We explored diverse, shifting ecologies (Star 1995) producing new knowledge collaboratively as scientists are perpetually (re)organizing.

The power of these coordinative entities lies in their ability to facilitate studies of the dynamics, and comparison, of how and when composite entities come together to function and support the practices that support scientific innovation. This study primarily investigated only the data collection, processing, and analysis work that was rooted in, and extending out of, four research groups. A different or larger set of questions, such as about how research results are distributed or where research questions come from or how citizen science functions, would likely yield still more types of arrangements. More research is needed to explore the similarities and differences that exist within entities, and within combinations of entities. A more appropriately nuanced way of understanding these organizational forms, and how and when they form and are invoked, can help us to much more effectively and appropriately support nascent science practices.

This typology is a move towards taking seriously the idea that CSCW can support more nuanced, yet not hopelessly complex, narratives about how and when scientists collaborate when undertaking data intensive research activities. We begin to find that the boundless tangle of scientific collaboration becomes a bit more legible. The collaborative space is still a complex muddle, but a little less so, and we inch closer to rendering it more tractable as design space. Our exposition of five coordinative entities opens the door for deeper analyses that shows not just that scientific collaborations have multiple forms at once or that they have permeable boundaries, but how they engage in relational work within and among diverse, dispersed forms of organizing.

8 Acknowledgments

The authors thank the anonymous reviewers and editors for their help refining this work. We thank Ying-Yu Chen and E. Illana Diamant for help with data collection, Erin Sy and Ron Piell for help with early analysis, and members of the Computer Supported Collaboration Laboratory including Andrew Neang, Michael Beach, Ridley Jones, Will Sutherland, and Os Keyes for feedback on drafts. We also wish to thank historian Dave Struthers for valuable criticism. Thanks also to Rebecca R., Charlie, Ash, and Ruby for support. Above all we are thankful to our research subjects for their time, insights, and feedback. Any errors are ours alone.

This work was supported by U.S. National Science Foundation grants IIS0954088 and ACI-1302272. Dr. Paine's work at Lawrence Berkeley National Laboratory is supported by the U.S. Department of Energy, Office of Science and Office of Advanced Scientific Computing Research (ASCR) under Contract No. DE-AC02-05CH11231. The views in this paper represent the authors and do not represent those of the U.S. National Science Foundation, Department of Energy, or the University of California.

References

Ackerman, Mark S.; Christine A. Halverson; Thomas Erickson; and Wendy A. Kellogg (2008).

Introduction. In Mark S. Ackerman, Christine A. Halverson, Thomas Erickson and Wendy A. Kellogg (eds), Resources, co-evolution and artifacts: Theory in CSCW. London, UK: Springer London, pp. 1–6.

Bates, Jo; Yu-Wei Lin; and Paula Goodale (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. Big Data & Society, vol. 3, no. 2, pp. 1–12.

Berman, F. (2001): The Human Side of Cyberinfrastructure. EnVision, vol. 17, no. 2, p. 1.

Bietz, Matthew J.; Eric P.S. Baumer; and Charlotte P. Lee (2010). Synergizing in Cyberinfrastructure development. Computer Supported Cooperative Work (CSCW), vol. 19, no. 3–4, pp. 245–281.

Bietz, Matthew J.; Toni Ferro; and Charlotte P. Lee (2012). Sustaining the development of Cyberinfrastructure: An organization adapting to change. In J. Grudin; G. Mark, and J.

Riedl (eds): CSCW'12. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, Seattle, Washington, USA. New York: ACM, pp. 901–910.

Bietz, Matthew J.; and Charlotte P. Lee (2009). Collaboration in Metagenomics: Sequence databases and the Organization of Scientific Work. In I. Wagner; H. Tellioglu; E. Balka; and C. Ciolfi (eds): ECSCW'09. Proceedings of the 11th European Conference on Computer Supported Cooperative Work, 7–11 September 2009. Vienna, Austria. London: Springer, pp. 243–262.

Birnholtz, Jeremy P.; and Matthew J. Bietz (2003). Data at work: Supporting sharing in science and engineering. In M. Tremaine and C. Simone (eds): GROUP'03. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, Florida, USA, 9–12 November 2003. New York: ACM, pp. 339–348.

Bonney, Rick; Caren B. Cooper; Janis Dickinson; Steve Kelling; Tina Phillips; Kenneth V. Rosenberg; and Jennifer Shirk (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. BioScience, vol. 59, no. 11, pp. 977–984.

Borgman, Christine L. (2007). Scholarship in the digital age: Information, infrastructure, and the internet. Cambridge, MA: MIT Press.

Borgman, Christine L. (2015). Big data, little data, no data: Scholarship in the networked world.

Cambridge, MA: MIT Press.

Bos, Nathan; Ann Zimmerman; Judith S. Olson; Jude Yew; Jason Yerkie; Erik Dahl; Daniel Cooney; Gary M. Olson (2008). From shared databases to communities of practice: A taxonomy of Collaboratories. In G. M. Olson; A. Zimmerman; and N. Bos (eds): Scientific collaboration on the internet. Cambridge, MA: MIT Press, pp. 53–72.

Bourdieu, Pierre (1984). Distinction: A social critique of the Judgement of taste. Cambridge, MA: Harvard University Press.

Bourdieu, Pierre; and Loïc J. D. Wacquant (1992). An invitation to reflexive sociology. Chicago, IL: Chicago University Press.

Bowker, Geoffrey C.; and Susan Leigh Star (1999). Sorting things out: Classification and its consequences. Cambridge, MA: MIT Press.

Charmaz, Kathy (2014). Constructing grounded theory: A practical guide through qualitative analysis. London, UK: Sage Publications.

Chompalov, Ivan; and Wesley Shrum (1999). Institutional collaboration in science: A typology of technological practice. Science, Technology, & Human Values, vol. 24, no. 3, pp. 338–372. Clarke, Adele E; and Susan Leigh Star (2008). The social worlds framework: A theory/methods package. In E. J. Hackett; O. Amsterdamska; M. Lynch; and J. Wajcman (eds): The handbook of science and technology studies. Cambridge, MA: MIT Press, pp. 113–137.

Cohn, Marisa Leavitt (2016). Convivial decay: Entangled lifetimes in a geriatric infrastructure. In P. Bjørn; and J. Konstan (eds): CSCW'16. Proceedings of the 19th ACM Conference

on ComputerSupported Cooperative Work & Social Computing, San Francisco, California, USA, 27 February 2 March 2016. New York: ACM, pp. 1511–1523.

Darch, Peter T., and Ashley E. Sands (2015). Beyond Big or Little Science: Understanding Data Lifecycles in Astronomy and the Deep Subseafloor Biosphere. In D. Bailey; T. Finholt (eds): 2015 iConference Proceedings, Newport Beach, CA, 24–27 March 2015. Grandville, MI: iSchools Inc. DiMaggio, Paul (1988). Interest and Agency in Institutional Theory. In Lynne G. Zucker (ed), Institutional patterns and organizations: Culture and environment. Cambridge, MA: Ballinger Publishing, pp. 3–21.

Edwards, Paul N. (2010). A vast machine: Computer models, climate data, and the politics of global. Cambridge, MA: MIT Press.

Edwards, Paul N.; Steven J. Jackson; Geoffrey C. Bowker; and Cory P. Knobel (2007). Understanding infrastructure: Dynamics, tensions, and design. University of Michigan, School of Information: Ann Arbor, MI.

Edwards, Paul N.; Matthew S. Mayernik; Archer L. Batcheller; Geoffrey C. Bowker; and Christine L. Borgman (2011). Science friction: Data, metadata, and collaboration. Social Studies of Science, vol. 41, no. 5, pp. 667–690.

Emerson, Robert M.; Rachel I. Fretz; and Linda L. Shaw (1995). Writing ethnographic Fieldnotes. Chicago, IL: The University of Chicago Press.

Faniel, Ixchel; and Trond Jacobsen (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work (CSCW), vol. 19, no. 3, pp. 355–375.

Finholt, Thomas (2002). Collaboratories. In B. Cronin (ed), Annual review of information science and technology. Washington, D.C.: American Society for Information Science, pp. 73–107.

Fligstein, Neil; and Doug McAdam (2012). A theory of fields. Oxford, UK: Oxford University Press. Fujimura, Joan H. (1987). Constructing `do-Able' problems in Cancer research: Articulating alignment. Social Studies of Science, vol. 17, no. 2, pp. 257–293.

Fujimura, Joan H. (1996). Crafting science: A Sociohistory of the quest for the genetics of Cancer. Cambridge, MA: Harvard University Press.

Galison, P.; and B.W. Hevly (1992). Big science: The growth of large-scale research. Palo Alto, CA: Stanford University Press.

Gergen, Kenneth J. (2010). Co-constitution, causality, and confluence: Organizing in a world without entities. In T. Hernes and S. Maitlis (eds): Process, Sensemaking, and organizing. Oxford, UK: Oxford University Press, pp. 55–69.

Gerson, Elihu M. (2008). Reach, bracket, and the limits of rationalized coordination: Some challenges for CSCW. In M. Ackerman; C. Halverson; T. Erickson; and W. Kellogg (eds): Resources, coevolution and artifacts: Theory in CSCW. London, UK: Springer, pp. 193–220.

Hanseth, Ole; and Nina Lundberg (2001). Designing work oriented infrastructures. Computer Supported Cooperative Work (CSCW), vol. 10, no. 3/4, pp. 347–372.

Harper, Richard H. R. (1997). Inside the Imf: An ethnography of documents, technology, and organizational action. New York, NY: Routledge.

Harper, Richard H. R. (2000). The organisation in ethnography - a discussion of ethnographic fieldwork programs in CSCW. Computer Supported Cooperative Work (CSCW), vol. 9, no. 2, pp. 239–264.

Howison, James; and James D. Herbsleb (2011). Scientific software production: Incentives and collaboration. In J. Bardram and N. Ducheneaut (eds): CSCW'11. Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, Hangzhou, China. New York: ACM, pp. 513–522.

Jirotka, Marina; Charlotte P. Lee; and Gary M. Olson (2013). Supporting scientific collaboration: Methods, tools and concepts. Computer Supported Cooperative Work (CSCW), vol. 22, no. 4–6, pp. 667–715.

Jirotka, Marina; Rob Procter; Tom Rodden; and Geoffrey Bowker (2006). Special issue: Collaboration in E-research. Computer Supported Cooperative Work (CSCW), vol. 15, no. 4, pp. 251–255.

Kaltenbrunner, Wolfgang (2017). Digital infrastructure for the humanities in Europe and the US: Governing scholarship through coordinated tool development. Computer Supported Cooperative Work (CSCW), vol. 26, no. 3, pp. 275–308.

Karasti, Helena; Karen S. Baker; and Florence Millerand (2010). Infrastructure time: Long-term matters in collaborative development. Computer Supported Cooperative Work (CSCW), vol. 19, no. 3–4, pp. 377–415.

Karasti, Helena; and Jeanette Blomberg (2018). Studying Infrastructuring ethnographically. Computer Supported Cooperative Work (CSCW), vol. 27, no. 2, pp. 233–265.

Kee, Kerk; and Larry Browning (2010). The dialectical tensions in the funding infrastructure of Cyberinfrastructure. Computer Supported Cooperative Work (CSCW), vol. 19, no. 3, pp. 283–308.

Knorr-Cetina, Karin (1999). Epistemic cultures: How the sciences make knowledge. Cambridge, MA: Harvard University Press.

Kraut, Robert; Carmen Egido; and Jolene Galegher (1988). Patterns of contact and communication in scientific research collaboration. In I. Greif (ed): CSCW'88. Proceedings of the 1988 ACM conference on Computer-Supported Cooperative Work, Portland, Oregon, USA, 26–28 September 1988. New York: ACM, pp. 1–12.

Kraut, Robert; Jolene Galegher; and Carmen Egido (1986). Relationships and tasks in scientific research collaborations. In I. Greif (ed): CSCW'86. Proceedings of the 1986 ACM conference on Computer-Supported Cooperative Work, Austin, Texas, 3–5 December 1986. New York: ACM, pp. 229–245.

Langhoff, Tue Odd; Mikkel Hvid Amstrup; Peter Mørck; and Pernille Bjørn (2018). Infrastructures for healthcare: From synergy to reverse synergy. Health Informatics

Journal, vol. 24, no. 1, pp. 43–53. Larkin, Brian (2013). The politics and poetics of infrastructure. Annual Review of Anthropology, vol. 42, no. 1, pp. 327–343.

Latour, Bruno (1987). Science in action. Cambridge, MA: Harvard University Press.

Latour, Bruno (2005). Reassembling the social. Oxford, UK: Oxford University Press.

Latour, Bruno; and Steve Woolgar (1986). Laboratory life: The construction of scientific facts. Princeton, NJ: Princeton University Press.

Lawrence, Katherine A. (2006). Walking the tightrope: The balancing acts of a large E-research project. Computer Supported Cooperative Work (CSCW), vol. 15, pp. 385–411.

Lee, Charlotte P. (2007). Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing Chaos in collaborative work. Computer Supported Cooperative Work (CSCW), vol. 16, no. 3, pp. 307–339.

Lee, Charlotte P.; Paul Dourish; and Gloria Mark (2006). The human infrastructure of Cyberinfrastructure. In P. Hinds and D. Martin (eds): CSCW'06. Proceedings of the 2006 20th anniversary conference on Computer Supported Cooperative Work, Banff, Alberta, Canada. New York: ACM, pp. 483–492.

Lee, Charlotte P., Drew Paine (2015). From the Matrix to a Model of Coordinated Action (MoCA): A Conceptual Framework of and for CSCW. In L. Ciolfi and D. McDonald (eds): CSCW'15. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, Vancouver, BC, Canada, 14–18 March 2015. ACM, pp. 179–194.

Leonelli, Sabina (2016). Data-centric biology: A philosophical study. Chicago, IL: University of Chicago Press.

Lynch, Michael (2002). Protocols, practices, and the reproduction of technique in molecular biology. The British Journal of Sociology, vol. 53, no. 2, pp. 203–220.

National Research Council (2015). Enhancing the effectiveness of team science. Washington, DC: The National Academies Press.

Newstead, Clare, Carolina K. Reid, Matthew Sparke (2003). The Cultural Geography of Scale. In K. Anderson; M. Domosh; S. Pile; and Nigel Thrift (eds): Handbook of Cultural Geography. London, UK: Sage Publications, pp. 485–497.

Olson, Gary M.; and Judith S. Olson (2000). Distance matters. Human-Computer Interaction, vol. 15, no. 2, pp. 139–178.

Olson, Gary M., Ann Zimmerman, and Nathan Bos (2008a). Scientific collaboration on the internet. Cambridge, MA: MIT Press.

Olson, Judith S., Erik C. Hofer, Nathan Bos, Ann Zimmerman, Gary M. Olson, Daniel Cooneyand Ixchel Faniel (2008b). A Theory of Remote Scientific Collaboration. In G. Olson, A. Zimmerman and N. Bos (eds): Scientific Collaboration on the Internet. Cambridge, MA: MIT Press, pp. 73–97.

Paine, Drew (2016). Software and Space: Investigating How a Cosmology Research Group Enacts Infrastructure by Producing Software. Ph.D. Dissertation. University of Washington: Dept. of Human Centered Design & Engineering, College of Engineering.

Paine, Drew, and Charlotte P. Lee (2014). Producing data, producing software: Developing a radio astronomy research infrastructure. In C. Medeiros (ed): eScience2014. IEEE 10th International Conference on e-Science, Guarujá, Brazil, 20–24 October 2014. New York: IEEE, pp. 231–238.

Paine, Drew, and Charlotte P. Lee (2017). "Who Has Plots?": Contextualizing Scientific Software, Practice, and Visualizations. Proceedings of the ACM on Human-Computer Interaction, vol. 1, no. CSCW, article 85.

Paine, Drew and Lavanya Ramakrishnan (2019). Surfacing Data Change in Scientific Work. In N. Taylor, C. Christian-Lamb, M. Martin and Bonnie Nardi (eds): iConference 2019. Information in Contemporary Society, Washington D.C., 31 March – 3 April 2019. Switzerland: Springer Cham, pp. 15–26.

Paine, Drew, Erin Sy, Ying-Yu Chen and Charlotte P. Lee (2014). Data, software, and advanced computational usage of University of Washington Research Leaders. University of Washington, Seattle, WA: Computer Supported Collaboration Laboratory, Dept. of Human Centered Design & Engineering, College of Engineering.

Paine, Drew, Erin Sy, Ron Piell, Charlotte P. Lee (2015). Examining Data Processing Work as Part of the Scientific Data Lifecycle: Comparing Practices across Four Scientific Research Groups. In D. Bailey, T. Finholt (eds): 2015 iConference Proceedings, Newport Beach, CA, 24–27 March 2015. Grandville, MI: iSchools Inc.

Plantin, Jean-Christophe (2019). Data cleaners for pristine datasets:Visibility and invisibility of data processors in social science. Science, Technology, & Human Values, vol. 44, no. 1, pp. 52–73.

Ribes, David (2014). The kernel of a research infrastructure. In M. Morris and M. Reddy (eds): CSCW'14. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, Baltimore, Maryland, USA, 15–19 February 2014. New York: ACM, pp. 574587.

Ribes, David (2017). Notes on the concept of data interoperability: Cases from an ecology of Aids research infrastructures. In L. Barkhuus; M. Borges; and W. Kellogg (eds): CSCW'17. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland, Oregon, USA, 25 February – 1 March 2017. New York: ACM, pp. 1514–1526.

Ribes, David; and Thomas A. Finholt (2009). The long now of technology infrastructure: Articulating tensions in development. Journal of the Association for Information Systems, vol. 10, no. 5, pp. 375–398.

Ribes, David; and Charlotte P. Lee (2010). Sociotechnical studies of Cyberinfrastructure and E-research: Current themes and future trajectories. Computer Supported Cooperative Work (CSCW), vol. 19, no. 3, pp. 231–244.

Rolland, Betsy and Charlotte P. Lee (2013). Beyond trust and reliability: Reusing data in collaborative Cancer epidemiology research. In C. Lampe and S. Counts (eds): CSCW'13. Proceedings of the 2013 conference on Computer Supported Cooperative Work, San Antonio, Texas, USA, 23–27 February 2013. New York: ACM, pp. 435–444.

Schmidt, Kjeld (1990). Analysis of cooperative work: A conceptual framework. Risø National Laboratory.

Schmidt, Kjeld; and Carla Simone (1996). Coordination mechanisms: Towards a Conceptual Foundation of CSCW systems design. Computer Supported Cooperative Work (CSCW), vol. 5, no. 2, pp. 155–200.

Schmidt, Kjeld; and Ina Wagner (2004). Ordering systems: Coordinative practices and artifacts in architectural design and planning. Computer Supported Cooperative Work (CSCW), vol. 13, pp. 349–408.

Star, Susan Leigh (1995). Introduction. In S. L. Star (ed), Ecologies of knowledge: Work and politics in science and technology. Albany, NY: State University of New York Press, Albany.

Star, Susan Leigh and Karen Ruhleder (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, vol. 7, no. 1, pp. 111–134. Star, Susan Leigh and Anselm Strauss (1999). Layers of silence, Arenas of Voice: The Ecology of Visible and Invisible Work. Computer Supported Cooperative Work (CSCW), vol. 8, no. 1-2, pp. 9–30.

Steinhardt, Stephanie B.; and Steven J. Jackson (2014). Reconciling rhythms: Plans and temporal alignment in collaborative scientific work. In M. Morris and M. Reddy (eds): CSCW'14. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, Baltimore, Maryland, USA, 15–19 February 2014. New York: ACM, pp. 134–145.

Strauss, Anselm (1988). The articulation of project work: An organizational process. The Sociological Quarterly, vol. 29, no. 2, pp. 163–178.

Velden, Theresa (2013). Explaining field differences in openness and sharing in scientific communities. In C. Lampe and S. Counts (eds): CSCW'13. Proceedings of the 2013 conference on Computer Supported Cooperative Work, San Antonio, Texas, USA, 23–27 February 2013. New York: ACM, pp. 445–458.

Vertesi, Janet; and Paul Dourish (2011). The value of data: Considering the context of production in data economies. In J. Bardram and N. Ducheneaut (eds): CSCW'11. Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, Hangzhou, China, 19–23 March 2011. New York: ACM, pp. 533–542.

Weiss, Robert Stuart (1995). Learning from strangers: The art and method of qualitative interview studies. New York, NY: The Free Press.

Wiggins, Andrea (2013). Free as in puppies: Compensating for Ict constraints in citizen science. In C. Lampe and S. Counts (eds): CSCW'13. Proceedings of the 2013 conference on Computer Supported Cooperative Work, San Antonio, Texas, USA, 23–27 February 2013. New York: ACM, pp. 1469–1480.

Wiggins, Andrea; and Kevin Crowston (2010). Developing a conceptual model of virtual organizations for citizen science. International Journal of Organisational Design and Engineering, vol. 1, no. 1&2, pp. 148–162.

Wulf, William A. (1993). The Collaboratory opportunity. Science, vol. 261, no. 5123, pp. 854–855.

Wyatt, Sally; and Brian Balmer (2007). Home on the range: What and where is the middle in science and technology studies? Science, Technology, & Human Values, vol. 32, no. 6, pp. 619–626.

Yasuoka, Mika (2009). Bridging and breakdowns - using computational artifacts across social worlds. Ph.D. dissertation. IT University of Copenhagen, Copenhagen, Denmark.

Yasuoka, Mika (2015). Collaboration across professional boundaries – The emergence of interpretation drift and the collective creation of project jargon. Computer Supported Cooperative Work (CSCW), vol. 24, no. 4, pp. 253–276.

Zimmerman, Ann S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology & Human Values, vol. 33, no. 5, pp. 631–652.